European Commission

Big data for monitoring educational systems

Education and Training

EUROPEAN COMMISSION

Directorate-General for Youth, Sport and Culture
Directorate A: Policy Strategy and Evaluation
Unit A.1: Strategy and Investments

Contact: Brigitte Devos

E-mail: Brigitte.Devos*@ec.europa.eu*

European Commission
B-1049 Brussels

EUROPEAN COMMISSION

Big Data for monitoring educational systems

Authors:

Bettina Berendt

Allison Littlejohn

Philippe Kern

Piotr Mitros

Xanthe Shacklock

Michael Blakemore

2017

Directorate-General for Education and Culture
Directorate A - Policy Strategy and Evaluation

EN

Disclaimer

This document has been prepared for the European Commission; however, it reflects the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

Acknowledgement

The authors are grateful to Geir Ottestad (European Commission) for his guidance and recommendations during the preparation of this document.

Luxembourg: Publications Office of the European Union, 2017

PDF ISBN 978-92-79-68693-1 doi: 10.2766/38557 NC-04-17-442-EN-N

Table of Contents

Acronyms

AI	Artificial Intelligence
ATO	Australian Tax Office
BDA	Big Data Analytics
BYOD	Bring your own device
CC	Creative Commons
CCTV	Closed-circuit television
CEDEFOP	European Centre for the Development of Vocational Training
CEF	Connecting Europe Facility
CMS	Classroom Management Software
CVET	Continuing Education and Training
DG EAC	Directorate General for Education and Culture
ECEC	Early childhood education and care
EC3	EUROPOL European Cybercrime Centre
ECTS	European Credit Transfer and Accumulation System
EDTECH	Education technology
ENISA	EU Agency for Network and Information Security
ERASMUS+	European Union programme for education, training, youth and sport
ESCS	Economic, social and cultural status
ESS	European Statistical System
ESSnet	Network of European Statistical System (ESS) organisations
ETER	European Tertiary Education Register
ET 2020	Education and Training Strategy 2020
ETwinning	EU platform for school staff to communicate, collaborate, develop projects
Europe 2020	Europe 2020 Strategy
EUROPOL	European Union's law enforcement agency
EURYDICE	Network on education systems and policies in Europe
FSRDC	Federal Statistical Research Data Centers (USA)
GDP	Gross Domestic Product
GDPR	General Data Protection Regulation
GPS	Global Positioning System
H2020	Horizon 2020 EU Research and Innovation programme
HEI	Higher Education Institution
ICT	Information Communication Technologies
ISCED	International Standard Classification of Education
ISCO	International Standard Classification of Occupations
K-12	The sum of primary and secondary education
MOOC	Massive open online course
NARIC	National Academic Recognition Information Centres
NEA	National education accounts
NIS	Network and Information Security
OECD	Organisation for Economic Co-operation and Development
OER	Open Educational Resources
OFSTED	Office for Standards in Education, Children's Services and Skills (UK)
OMC	Open Method of Coordination
PET	Privacy enhancing technologies
PIA	Privacy Impact Assessment
PIRLS	Progress in International Reading Literacy Study
PISA	Programme for International Student Assessment
RFID	Radio frequency identification
SIC	Safer Internet Centres
SMS	School Management Software
TIMMS	Trends in International Mathematics and Science Study

U-Multirank	A user-driven, multidimensional, world ranking of universities and colleges
UNESCO	United Nations Educational, Scientific and Cultural Organization
UNICEF	United Nations International Children's Emergency Fund
UOE	UNESCO/OECD/EUROSTAT database on education statistics
VET	Vocational Education and Training
VLE	Virtual Learning Environment
WCAG	Web Content Accessibility Guidelines

Executive summary

This report considers "*how advances in big data are likely to transform the context and methodology of monitoring educational systems within a long-term perspective (10-30 years) and impact the evidence based policy development in the sector*", big data are "*large amounts of different types of data produced with high velocity from a high number of various types of sources.*" Five independent experts were commissioned by Ecorys, responding to themes of: **students' privacy**, **educational equity and efficiency**, **student tracking**, **assessment** and **skills**. The experts were asked to consider the "*macro perspective on governance on educational systems at all levels from primary, secondary education and tertiary – the latter covering all aspects of tertiary from further, to higher, and to VET*", prioritising primary and secondary levels of education.

The contributions connect firmly to the EU policy context concerning the modernisation and digitalisation of education systems. The Europe 2020 strategy's flagship initiatives "*Digital Agenda for Europe*" and "*An Agenda for New Skills and Jobs*" align the changing societal and economic context for education, and the "*Digital Single Market Strategy*" calls for a strengthening of skills useful for the digital world. The strategic objectives in the ET 2020 Framework for EU cooperation in Education and training point to improving both the quality of education and training and also to promote equity, social cohesion and active citizenship. Big data can be considered as an accelerator of these policy goals through more effective monitoring. However, there are important considerations that need to be addressed in the development of advanced monitoring systems.

Most current data sources used for monitoring provide limited longitudinal insights, relying mostly on comparisons between time periods. For example, data on expenditure, register data, national, regional or local data on student performance, or international comparable data from Eurostat at the EU level, and international assessments such as the OECD PISA programme for reading, mathematics and science. While showing change over time, and across geographical units, such sources do not provide clear insights into the rate of change. Data can suffer from a 'collection to publication lag': they are collected at or around a particular date, and often require substantial post-processing before they can be analysed and results published. Data when published relate to the past, but as data 'ages' through time, it often has residual authority well beyond its 'temporal decay': it is the 'latest available' data and is often used as a key reference.

This presents challenges as the governance of education systems becomes more complex, and as more actors and stakeholders (students, teachers and managers, politicians, interest groups, researchers etc.) are involved. Some systems have de-centralised power away from a single national structure, sometimes introducing market forces. This makes the challenge of monitoring substantial, in particular as existing monitoring systems rely more on combining data from multiple sources, over different time periods, at varying levels of aggregation, and using different methodologies.

Big data are produced in **real time**, often **constantly** (such as location data on smartphones or keystroke and click-navigation data), and at **individual levels** (student assessment on learning platforms, or embedding information from social media into education monitoring processes). Real-time data, consistently captured at the individual level, can be **rapidly aggregated** for monitoring at the education system level. Highly detailed individual data have both opportunities for personalisation of learning, as well as threats to privacy, or to unfair categorisation of students when analytic algorithms are used to analyse student data. To address the challenges of big data, the Commission 'Big Data Strategy' was launched in 2014, focusing primarily on the Digital Single Market. Importantly, the European **General Data Protection Regulation (GDPR)** sets foundations for privacy protection online, and the Commission has also supported the development of privacy-enhancing technologies (supporting data minimisation and

anonymisation), investigated big data in official statistical systems, and privacy and data protection implications of cloud computing (where student data may be held beyond national borders in commercial learning platforms).

In the first section of this study, on **students' privacy**, a core argument is that the GDPR provides a fundamental framework for big data developments in the monitoring of education systems. Developments will need to consider the values and laws that enable societies to (continue to) function on the basis of fundamental rights and democracy. The section emphasises that the potential significant expansion in coverage and resolution enabled by a big data approaches does not imply that 'more data mean better monitoring'. Challenges relate to data quality, measurement systems, how data can be effectively anonymised, whether long-term coherent data for learners in life-long 'eportfolios' risks pervasive surveillance, what types of big data can be used for 'effective' monitoring, who owns the data about students, what are the competences of the education institutions and their staff in analysing and interpreting big data, and whether a reliance on automation and analytics means that education pathways may become determined by private actors that are not subject to democratic control. The section strongly advises that data protection and privacy must be designed into the big data systems and processes. It draws a 'line in the sand' for the fundamentals of data protection and privacy to be designed into big data learning systems to be used in the EU. If the learning platforms and the analytics have data protection incorporated into their design, then national education systems may be in a better position to use big data to deliver more equity and efficiency, through individualised monitoring of assessment, better tracking for grouping students, and for this to be undertaken life-long (skills).

The section on **equity and efficiency** looks first at how existing data sources can be made 'big' through integration and interoperability across multiple data sources. Specific big data developments are presented, first for Portugal where the Troika required the development of a new education monitoring system. Second, a fully operational big data approach is shown for Estonia, integrating big data from student level to system level. Students and parents can access learning progress (transparency of process and openness of data), and (acknowledging the arguments in the privacy section) underpinning trust and confidence are powerful security and privacy protections applied to the national identity card, with very strong system security and cyber-defences. The section considers how individual student monitoring would not just look at their educational performance through learning platforms, but also give attention to value adding, personalised and blended learning. Student monitoring could flag learning issues by relating them to data from other relevant data systems, and assess whether the learning issues are purely educational, or are influenced by other social or external issues. It could change the nature of what is a 'school', for example overcoming the friction of location and distance to link the equity needs of students by resources delivered through learning platforms.

The section on **assessment** notes that big data cover a broad range of activities, such as on-line student social interactions, text, audio, and video data, and fine-grained interactions. In general, not dependent on but augmented by digitalization, student learning has moved towards a more cognitively active process, with rapid feedback, paths for remediation, and the ability for students to self-pace often resulting in learning gains. However, to date delivery of such gains has required costly investment in human capital (teachers etc.) that is located in traditional education institutions. A possibly more cost-effective approach is through learning at scale: having substantial course resources (increasingly, entire courses) shared by thousands of students. However, data from online learning platforms are mostly being gathered by the for-profit corporations who create the platforms. Educational data are considered proprietary, and since such data are divided among hundreds of educational technology corporations, there is no easy way to combine or correlate such data.

Furthermore, the volume and scale of data can require major investment to store and effectively analyse them. Data from at-scale platforms is currently in terabytes. Once multimedia data are included (for example, student conversations over video), it moves into the petabyte or exabytes scale. Analysing data rapidly and effectively is crucial, but post-hoc analysis currently mostly drives policy choices. Semester-to-semester feedback can lead to improved course design. Daily feedback can help identify where students are struggling, and provide feedback. Immediate feedback allows students to help them identify problem areas, and remedy knowledge gaps and misconceptions. However, processing data in real-time at the velocity coming out of at-scale learning systems, still has significant computational and competence related challenges.

The section looks ahead, considering whether educational resource production, data collection, and educational technology may become more unified and centralised, and if blended learning will be the norm. The role of the teacher is shifting from the primary source of information to working with students 1:1 utilising such digital materials. The trend today is that governments invest heavily in digitizing their education systems, but often lacking clear goals, guidance or impact assessment. In many ways, the landscape of educational technology resembles that of computing circa 1975, or e-commerce circa 1999; there are many competing platforms, and it is too early to tell which ones will dominate. But, without appropriate regulation and if a monopoly is in place, progress generally stops. Hence, emerging **policy challenges** for the EU can be found in the areas of data rights, privacy, security, research access and analytics, inclusion and equity, and (with increasing internationalisation of platforms) cross-border regulation.

The section on **student tracking** concerns the process of grouping students by ability. Understanding student behaviour and problems with learning can feed back into policy decisions about where and when students should learn specific skills. For example, groups of students with different abilities could be tracked and streamed into different educational trajectories based on assessments of their skill levels and predictions of future labour market needs. A key to the success of this vision is the measurement and use of appropriate big data, but also a thorough policy framework that secures the individual freedom of choice and rights to privacy. Using big data and algorithms to suggest paths for students, could tackle some tracking challenges. The section reviews challenges associated with the sourcing, storing and analysis of big data. Data sources are often distributed across multiple sources and servers, and are gathered using different methods. To enable comparability, data must have standardised quantitative and qualitative indicators that offer insights policy and practice. However, currently available data are mostly limited to online learner activity, and do not offer direct evidence of offline activity and cognitive and non-cognitive development. Data are often held in 'silos' at difference stages of lifelong learning.

Therefore, a major challenge for education systems in Europe is to implement student tracking in ways that enable us to extract meaning from large datasets being generated through micro-level, online student activity, and to distil this data into usable information for students, teachers, and governments. The policy questions at the EU level will need to address whether student tracking will promote equity, and whether 'success' can be more richly assessed with a wider range of data.

The section on **skills forecasting** moves firmly into the lifelong learning context. This contribution first identifies ways in which big data can be implemented in the analysis of labour market demands. Second, it outlines possible avenues of using educational big data in helping develop students' skills and to improve the responsiveness of educational systems to labour market skills demand. Finally, the opportunities and challenges that are discussed are taken into possible actions at EU level. Big data and analytics can help skills development across educational pathways which can help increase students' academic performances and help them make personalised career choices, matching better their education outcomes, skills, competencies, and labour market needs.

However, developments to date in linking skills, labour market, and education are limited. Big data are used in job vacancy and recruitments systems, for example through monitoring of the social media activities of applicants. There are opportunities for better monitoring of the skill needs of labour markets and linking it to more responsive and data-rich monitoring of education systems.

Policy recommendations on skills focus on the creation of a robust infrastructure and methodological framework enabling data collection and data analysis, and regulatory frameworks for privacy and governance of big data for education and skills. This can be undertaken through collaboration between labour market bodies and social partners, encouraging the digitization of the job vacancies which involve lower qualifications and vocational skills, integrating big data analysis as part of the Cedefop and Eurostat work on skills forecasting. The section finally looks ahead, noting the potential for significant diversification and fragmentation of education structures and the need for governments to work on supporting mechanisms to build big data across fragmented structures.

The contributors consider possible developments in a 10 to 30-year horizon, taking as a starting point that there will be fundamental changes in the relationships between teachers, students, employers, and families, emphasising the need for all parties to be part of the lifelong learning process. Curricula will need to break across subject boundaries, as will assessment. In 2017, we are in an environment of educational mobility, education and training transparency tools, eCommerce, eGovernment, remote working, wireless internet, robotics, social networks, the cloud, civilian access to GPS location, and disruptive business models such as Uber and Airbnb. Yet, while radical disruption has been evident in many sectors, radical organisational change has not been as evident in education systems. The focus of compulsory education at primary and secondary levels remains strongly oriented to formally structured school years based on physical presence at school, with nationally or federally specified curricula, and on the same age-based transitions from primary to secondary to tertiary education. One feature of education systems in the age of big data is that while they may have 'modernised' their teaching and learning within education layers, the borders between the layers are still very age-related, rather than lifelong learning oriented. That is not a particular concern if the **data discontinuities that exist across the borders of education levels** can be overcome, and big data shows considerable potential.

There could be two potentially converging approaches to fully modernising education systems. First, **restructuring and organisational change** can continue to happen, although this will need to engage with actors such as teacher unions as teaching and learning patterns change. Paradoxically, the education systems which are experiencing the most challenges over quality and labour market relevance could be the ones most suitable for change – their legacy effects will be less than those more advanced, but still conventionally structured, systems.

Second, **big data integration and transparency tools could enable learners to 'transport' their learning across all the layers of education, from primary to life-long learning.** For example, from formal to informal, and from education that leads to a formal qualification, to a collection of recognised digital badges. The latter process would be strongly reliant on the use of big data, ePortfolios, and more flexible systems of recognition (supported by robust classification systems for skills, competencies, qualifications etc.). To achieve this big data analytics and developments in artificial intelligence will be important. And, quite apart from the need to improve the efficiency and effectiveness of education systems, the business potential in developing learning and monitoring systems is significant, since governments spent about $3 trillion globally on education annually. There may be very different partnership arrangements between governments, their education systems, and commercial providers of education services.

Introduction and summary

Aims and methodology

This document is the outcome of a request from DG EAC to consider:

> "*how advances in big data are likely to transform the context and methodology of monitoring educational systems within a long-term perspective (10-30 years) and impact the evidence based policy development in the sector*".

The Commission definition of big data concerns "*large amounts of different types of data produced with high velocity from a high number of various types of sources*" (Commission, 2014d). More extensive data, collected and analysed more rapidly, and cost-effectively, has the potential to better inform the Commission and Member States in monitoring their progress towards (and beyond) the objectives of the Europe 2020 Strategy and beyond.[1]

Five independent experts were commissioned by Ecorys UK to respond to the request in the context of the five themes related to the monitoring of education systems: students[2] privacy, educational equity and efficiency, student tracking, assessment, and skills (See Annex A for details). They were asked to consider the "*macro perspective on governance on educational systems at all levels from primary, secondary education and tertiary – the latter covering all aspects of tertiary from further, to higher, and to VET*", with a clear priority being on primary and secondary levels of education.

The experts (biographical information in Annex B) were each asked to cover one of the themes:

- Privacy of students (Bettina Berendt);
- Educational efficiency and equity (Xanthe Shacklock);
- Assessment (Piotr Mitros);
- Student tracking (Allison Littlejohn);
- Skills forecasting (Philippe Kern).

Michael Blakemore (Ecorys) coordinated this introduction and summary section.

Education, systems, and monitoring

At its most basic an education system can be comprised of four elements: teachers, students, a context, and content. Conceptually speaking, a single child (student) being taught (by a parent, as an educator, who is delivering content), in a home environment (context) could be regarded as an education system. In reality, education predominantly has taken place in an institutional environment (schools, universities, colleges etc.) organised along formal lines determined by government - the resulting education system is complex.

The Eurydice network observes that education systems are multi-level (local, regional, federal, national), that they educate increasingly diverse societies (multi-cultural, ageing) which have different "*values and identities*", that stakeholders are diverse (parents, students, policymakers, teacher unions, commercial education providers, researchers, think-tanks) and that there are very strong views across stakeholders about what education systems should provide to students, how they should be structured, what content should be in curricula, what should be the relationship between teaching and learning and the world of work, or how assessment should be undertaken (Eurydice, 2017b).

[1] http://ec.europa.eu/europe2020/europe-2020-in-a-nutshell/index_en.htm
[2] "Students" is being used in this document as a generic term to cover learners, pupils, students etc.

In its description of education systems[3] Eurydice specifies 14 categories that elaborate beyond the four elements:

- Political, Social and Economic Background and Trends;
- Organisation and Governance;
- Funding in Education;
- Early Childhood Education and Care;
- Primary Education;
- Secondary and post-Secondary non-Tertiary Education;
- Higher Education;
- Adult Education and Training (which would include professional learning and training and lifelong learning);
- Teachers and Education Staff;
- Management and Other Education Staff;
- Quality Assurance;
- Educational Support and Guidance;
- Mobility and Internationalisation;
- Ongoing Reforms and Policy Developments.

While the Eurydice categorisation details the structure of education systems, what happens in the context of students and their 'learning' is complex. Learning can be formal, informal, or non-formal, and the EU has been active in promoting the validation and recognition of all types of learning (CoR, 2014). Frick elaborates further, for example noting for example, that learning can be accidental, the result of individual discovery, it can be the result of "*disciplined enquiry*" where teachers and students work together, or students work together (Frick, 2016).

Across the components of an education system there are multiple creators and users of information relevant for monitoring. Budgets are set by government, and distributed through the system via structures such as local government and school districts. School governors can be provided with statistics gathered for them, sometimes on integrated 'dashboards' with statistics that tend to be those provided to them by education authorities – these are often data that have already been provided by their schools.

Teachers, and their schools collect prolific amounts of data relating to student attainment and progress, and the 'quality' of the education of their students, and the outcomes of education, have become a major focus of policy makers. To date attainment has been predominantly measured through processes such as standardised tests and through formal reviews of school standards by external quality agencies. In the USA, a student progressing from pre-school to K-12 grade will have completed "*112 mandated standardized tests*", but (see the following discussion about PISA) this level of monitoring has not led to the US public school system performing better than countries with significantly fewer tests (Layton, 2015).

There are negative feedback situations where what can be excessive levels of monitoring can divert teacher resources away from the most important task of teaching and learning, for "*many teachers are overwhelmed by poor-quality data-collection*" (ECONOMIST, 2016g). The picture to date is of an incremental process where more data

[3] https://webgate.ec.europa.eu/fpfis/mwikis/eurydice/index.php/Countries

are being produced for monitoring purposes, resulting in data-production overload on teachers and managers, with a resulting heterogeneous landscape of metrics and indicators. Unevenness also exists as the result of situations where the data captured and analysed sometimes are not the data needed, and which may drive the wrong sorts of behaviours. For example, using league tables of assessment can result in narrowing the teaching activities to focus on position in the table. Therefore, data needed to measure learning are difficult to capture and analyse because they tend to be contextualised.

This has concerned the OECD which advises that "*the use of evaluation and assessment results should avoid distortions in the education process such as teaching-to-the-test and narrowing of the curriculum*" (OECD, 2013). This study explores how 'big data' may have the potential to provide more efficient (less administrative burden) and effective (more timely, accurate, and informative indicators) monitoring. To date many education data sources (covering learners, researchers, organisations, teaching quality etc.) and are representative samples or complete surveys. They can be rich in detail and coverage, but are time-bound (published long after data collection), particularly when collected on an international level (such as TIMMS and PIRLS[4], the OECD PISA[5], or Eurostat at national[6] and sub-national levels, such as the 2016 Regional Yearbook[7]), since the tasks of data validation, harmonisation, and publication have to be undertaken with care.

Most current data sources used for monitoring provide limited longitudinal insights into the themes covered, relying very much on comparisons between time periods. While they can show that change occurs over time, and across geographical units (primarily country to regional to local administrative and governance geographies), they do not provide clear insights into the rate of change. For example, they will not clearly show whether there has been incremental change, or whether change was sudden (perhaps as the result of a particular policy action) and occurred at one particular stage in the times between data collection. As the terms of reference for this study noted:

> "*both national and international surveys are suffering from high costs, survey fatigue, a risk of sub-standard representativeness due to complex sampling procedures and a rather large component of self-reported data rather than measurements of actual practices*".

Such data series suffer from a 'collection to publication lag'. They are collected at or around a particular date, and then often require substantial post-processing before they can be analysed and results published. Data inevitably relate to times in the past. As the data 'age' through time, however, they often have residual authority well beyond their 'temporal decay', since they are the 'latest available' data and are used as key reference points. PISA statistics for 2012, covering 65 economies[8], were heavily cited in studies until the release of PISA 2015, covering 72 economies, and which was published in December 2016 (OECD, 2016d). Consequently, data are often not well-suited to monitoring needs, and a recent Joint Research Centre report, looking at the assessment of education equity, concluded that data problems continue to hinder monitoring (Hippe et al., 2016).

The OECD observes that governance of systems has become more complex as more actors and stakeholders (students, school teachers and managers, politicians, interest groups, researchers etc.) become involved (Burns and Köster, 2016). For example, some systems have de-centralised power away from a single national structure, sometimes introducing market forces, and the OECD study considers the question of what

[4] http://timss.bc.edu/
[5] https://www.oecd.org/pisa/
[6] http://ec.europa.eu/eurostat/web/education-and-training/statistics-illustrated
[7] http://ec.europa.eu/eurostat/documents/2995521/7627725/1-14092016-BP-EN.pdf/ce9d6051-d4d4-49e4-8803-236621150d96
[8] https://www.oecd.org/pisa/aboutpisa/

governance innovations are needed to cope with the diversity of actors (Burns and Köster, 2016). With such complexity, and heterogeneity, the challenge of monitoring is substantial, but existing monitoring systems rely more on combining data obtained from multiple sources, over different time periods, at varying levels of aggregation, and using different methodologies.

The variety of monitoring mechanisms that enable comparative assessment across national systems is shown in detail by the Eurydice reports on "*Structural indicators for monitoring education and training systems in Europe*" (Eurydice, 2015a). For example, in the area of schools (what it terms as monitoring 'achievement in basic skills') Eurydice lists four areas:

- Nationally standardised tests in literacy, mathematics and science;
- Recent national reports on achievement in the basic skills;
- Use of student performance data in external school evaluation;
- Central guidelines on addressing low achievement in initial teacher education (ITE).

There are internationally standardised tests, such as the OECD PISA[9] tests on science, reading and mathematics, which are at the current forefront of international comparative monitoring of some key elements of education systems at the school level.

PISA is an objective measure, but it is logistically challenging to undertake. Its content and structure needs political agreement across all the education systems that decide to participate. The education systems must plan the undertaking of the tests, and results need to be independently assessed, analysed, verified, and harmonised. The results have political impact, since the 'league list' position in PISA is often all that is reported. For example, the OECD press release for PISA 2015 states "*Singapore outperforms the rest of the world*" (OECD, 2016e), and much press coverage focuses on league list position (places 'lost' or 'gained'), even if a country may have actually increased the overall PISA scores, though also moved down a rank list. It is an indicator that can have political ramifications well beyond its intention.

Furthermore, comparing advanced developed countries with more developing ones is challenging. A country may be low on the scores, but have delivered significant improvements from an initial low level, and "*to know how effective an education system really is, we need to know where children are when they enter school and what progress the schools are responsible for. To start with we need a baseline*" (Tymms, 2016).

PISA data are problematic where there is increasing heterogeneity in a population. It is argued that part of the reason countries like Sweden have dropped in the tables is probably because of the increasing heterogeneity of the population, and the impact of migrant children learning in a second or third language. However, the OECD responded that the overall number of migrant children was not in itself enough to cause the reduction, and that "*we see that for students with less resources at home in terms of parents who help with their homework, the gap between them and other students is increasing*" (Roden, 2016).

PISA measures absolute levels of attainment, not value-adding. It is a surrogate measure relating to complex processes, influenced for example by the levels of training and professional development of teachers, the teaching environment (IT resources, curriculum, and pedagogy). Outcomes can be affected by the extent to which students with special needs (cognitive, disability) are specifically addressed.

[9] https://www.compareyourcountry.org/pisa?lg=en

At the higher education level, monitoring is more developed through international comparative monitoring. Nevertheless, monitoring still has undesired consequences in terms of the behaviours triggered by data that are surrogate measures. For example, the UK National Student Satisfaction Survey[10] has assessed 'student satisfaction' with their teaching and learning and assessment experience. The 2017 survey[11] is primarily oriented to final year undergraduates, and asks questions such as: "*Staff are good at explaining things; Staff have made the subject interesting; The course is intellectually stimulating; My course has challenged me to achieve my best work*". The data therefore represent a 'snapshot' of opinion by different cohorts, without a longitudinal dimension (for example a second survey one year after graduation for a more reflective opinion). Furthermore, this type of survey, used as a surrogate measure of teaching quality, is undertaken by students who are 'adults'. Such opinion surveys could be challenging to operate at school levels with very young students.

At the European level, the U-Multirank[12] system measures higher education institutions (HEIs) across dimensions of: teaching and learning; research; knowledge transfer; international orientation; and regional engagement. The HEInnovate[13] tool assesses HEIs across: leadership and governance; organisational capacity: funding, people and incentives; entrepreneurial teaching and learning; preparing and supporting entrepreneurs; knowledge exchange and collaboration; and, the extent to which an HEI is 'internationalised'.

There are well-developed global monitoring indicators that have significant impact on the reputations of HEIs, in particular QS Global Rankings[14], which show that the more indicators become global the less information can be robustly harmonised so that indicators are comparable. The six indicators, with their weights, are: academic reputation (40%); employer reputation (10%); student-to-faculty ratio (20%); citations per faculty (20%); international faculty ratio (5%) and international student ratio (5%). Such HEI indicators do not measure directly factors such as research excellence, or teaching quality, learning or subsequent success measures in employability, or career trajectory. Instead, surrogate measures are used, such as journal citations as a proxy for research quality, or student opinions of their teaching experience as a substitute for teaching quality. Furthermore, any weightings applied to each of the six indicators will also determine the outcomes, and the EU U-Multirank system particularly avoids the judgmental risk of weights (Ziegele and van Vught, 2017).

There have been significant developments in big data and the monitoring of education in HEIs. Universities are often self-standing institutions, and therefore have much more control over their information structures, although big data developments in HEIs that enable comparative longitudinal monitoring across HEIs are not yet developed. At national levels, quality assurance agencies impose their own metrics which universities then respond to, having to cope with the additional administrative burden. Nevertheless, the HEIs international comparative metrics, with all their inconsistencies, remain highly influential monitoring tools. Furthermore, there is no guarantee that big data and big data analytics implicitly lead to more efficiency. For example, Google Scholar Metrics[15] provide dynamic insights research at thematic, journal, and individual levels, potentially adding more burden for researchers when deciding on the publication channels for their research.

[10] http://www.hefce.ac.uk/lt/nss/
[11] http://www.thestudentsurvey.com/about.php
[12] http://www.umultirank.org
[13] https://heinnovate.eu/
[14] http://www.topuniversities.com/university-rankings
[15] https://scholar.google.es/citations?view_op=top_venues&hl=en&vq=eng_datamininganalysis

This study is tasked with looking particularly at the potential for big data developments at the primary and secondary levels, where UNESCO concludes that the monitoring and evaluation of most education systems lacks *"precision, effectiveness and efficiency", especially when the outcomes of education are to be measured*" (UNESCO, 2016a). The OECD further observes that monitoring and evaluation can only realistically be effective where an integrated approach is taken, and that the component parts (assessment, teacher training and appraisal, management and leadership evaluations etc.) should be monitored in an integrated manner (OECD, 2013).

Furthermore, the OECD emphasises that monitoring needs to have a much stronger longitudinal ability, and that there is the potential for big data approaches to overcome the problems of having to rely on comparisons over time periods without being able to fully understand the ways in which change occurred: "*longitudinal information systems could lead to a new culture of individual, organisational and sectorial learning in education, and thus of continuous improvement and innovation*" (Vincent-Lancrin and González-Sancho, 2014).

Big data monitoring of the achievement levels of learners (OECD, 2015c) can lead to the availability of 'hyperpersonal' learning where the classroom is used not to deliver homogenised lessons, but to provide focused individual learning that maximises the value-adding activity in learning (ECONOMIST, 2016b). However, there are logistical challenges when learning is not limited to a single platform, and when gathering data across platforms is required.

Furthermore, while the potential for improvement in monitoring is evident, the caution also is that 'more data is not necessarily better'. The OECD cautions that improved monitoring of big data needs to be set against some significant challenges. These include concerns over who owns data, how it is protected in IT systems, which privacy regulation is needed, what ethical guidelines should be agreed, what legislation is needed to protect data relating to students and staff, and what governance changes will be needed.

Education system change partly depends on political will, resourcing, and capacity of an education system to 'absorb' new developments. This may require changes in the cultural values associated with education, such as viewing education as lifelong, and not limited to the activities just within formal educational institutions. A clear example of how an education system can be integrated from individual to national levels is in Estonia, the highly successful eGovernment services have built strong trust among citizens in protecting their data while delivering sophisticated and efficient electronic services. The Estonia eSchool[16] service is explored in the sections on privacy, equity and efficiency.

Another example is the SIGA system in Portugal, where the involvement of the Troika during the economic crisis resulted in a request to the Portuguese Government to develop an evidence-base that would enable effective monitoring of "*the efficiency and efficacy of education*" (Evaristo, 2014). The SIGA system is also explored in the section on equity and efficiency.

With political commitment, and economic need, a national-level big data approach can be developed, but there is a natural caution by official statisticians in the adoption of new data and analytical methodologies. A report on skills measurement by the European Statistical Office (Eurostat) observes that while big data developments are evident in business, "*its use for official statistics still needs to be thoroughly assessed*" (ESTAT, 2016c).

Big data innovation in monitoring education systems is also set within the context of an increasingly monitored and interconnected society, driven by rapid developments in areas such as artificial intelligence (AI), the emergence of 'smart cities', and of changes

[16] https://e-estonia.com/component/e-school/

in how public services are designed and delivered (Eurofound, 2016). In that context, looking ahead 10-30 years, the monitoring of education systems is an activity needing to balance objectivity against the risk of fantasy.

The speed of technological development is intimidating, as communicated by the Global Challenges Foundation which looked ahead at major future risks to human civilisation. In the context of some of their forecasts, could functions of teachers even be replaced by AI through the what are termed "*whole brain emulations"* which are created after scanning and digitising human brains or considering that "*technology, political and social change may enable the construction of new forms of governance*" (Pamlin and Armstrong, 2015).

Could education even become a continuous experiment where continuous experimentation with education and schools becomes part of a computational analytical process of governance (Williamson, 2015b)? And, if smart cities are to develop equitably, how can policy makers avoid a widening of technological disparities experienced between urban and rural areas (rural areas often have worse internet infrastructure), and between advantaged and disadvantaged students, since:

> "*In 21 out of 42 countries and economies, disadvantaged students spent more time online than advantaged students. In all countries/economies, what students do with computers, from using e-mail to reading news on the Internet, is related to students' socio-economic status. Advantaged students are more likely than disadvantaged students to search for information or read news on line. Disadvantaged students, on the other hand, tend to use the Internet to chat or play videogames at least as often as advantaged students do.*" (OECD, 2016a)

The EU big data context

With the dramatic increase in data production come challenges to policy. Speed and volume are characteristics of big data, have been used effectively in electoral targeting during the 2016 US Presidential election (Wood, 2016), there are concerns that democratic engagement can be compromised when information is exchanged at such speeds that little time is available for reflection and debate (Bartlett and Grabbe, 2015). With openness of systems and data for online learning come additional challenges of system security and the threat of cybercrime (ESTAT, 2016a).

Other issues include who 'owns' data produced across multiple sources (such as blogs), whether it can be used in an education monitoring context, how can privacy and identity be protected (EDPS, 2016), and whether it is possible to completely anonymise data in a way that avoids identity being retrospectively known (the section on assessment considers this issue). In that context, the Commission is exploring how big data anonymisation could be achieved in healthcare services using blockchain technology (Commission, 2017d).

There are considerations about what rights do individuals have to know about or challenge analytical decisions which affect them, even being given the right to have an explanation about algorithms used to analyse their educational progression (ECONOMIST, 2016e). However, it is important that criticisms of algorithms are not targeted just on big data: teachers have been applying 'algorithms' for decades, and they are often individual and idiosyncratic marking and judgment criteria that they use when assessing student work (Gee, 2015).[17]

In March 2017, the European Parliament agreed a resolution on the "*fundamental rights implications of big data*" (EP, 2017), emphasising the need for strong privacy, and that big data analytics should not discriminate between people, and highlighting also the need

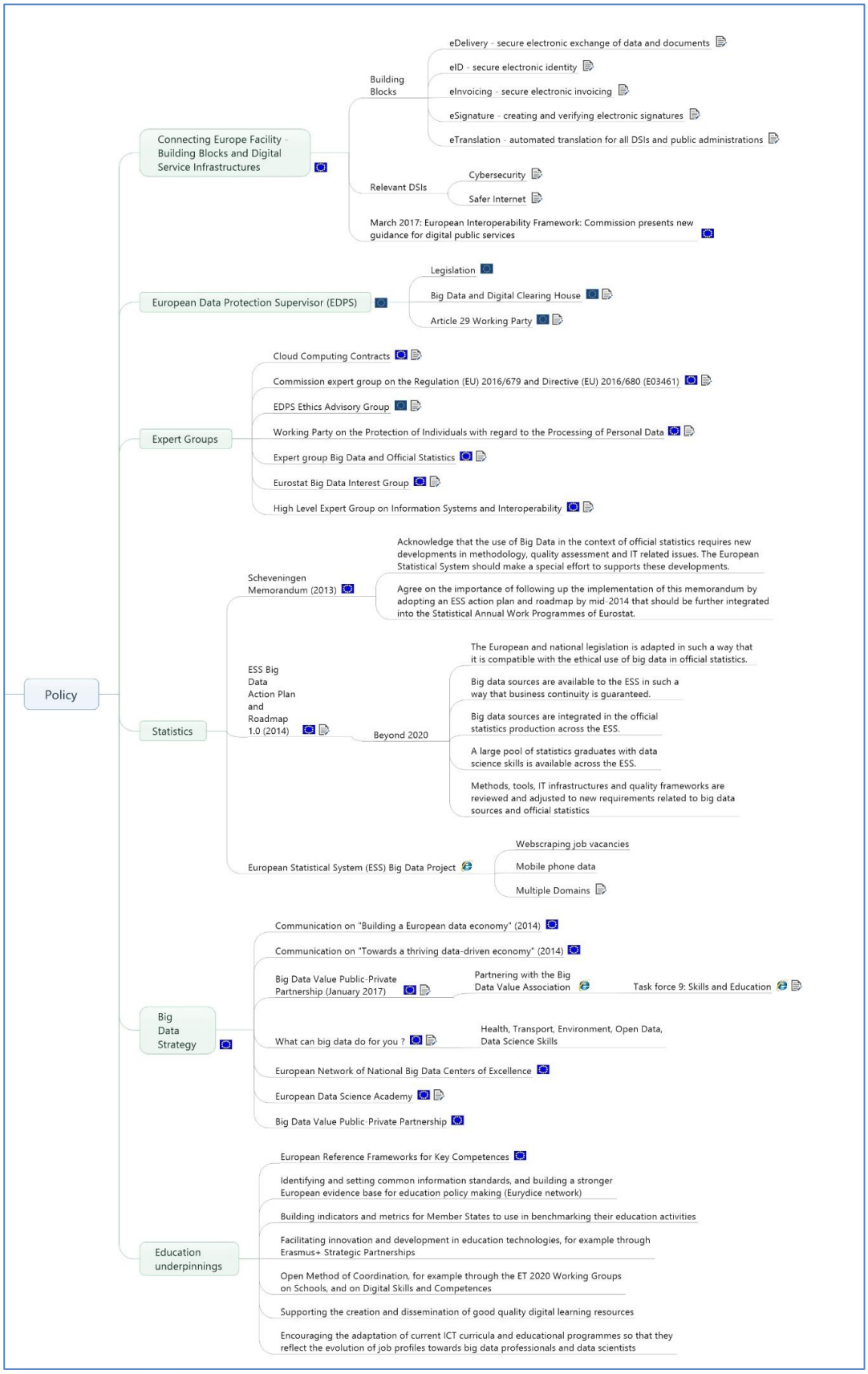
[17] This is being addressed by the Horizon 2020 funded TESLA Project – "*an adaptive trust e-assessment system for assuring e-assessment processes in online and blended environments*" http://tesla-project.eu/

for "*algorithmic accountability and transparency*" (Monteleone, 2017). As the European Data Protection Supervisor observes, there are concerns if algorithms are regarded as business property and are known only by the commercial owners (Buttarelli, 2016).

The Joint Committee of the European Supervisory Authorities has undertaken a consultation on big data and the financial profiling of customers, emphasising that the algorithms that are used in big data analytics must be shown to be unbiased, otherwise the benefits of analysis will be diminished (ESMA, 2016). The Commission strongly emphasises the foundations of data protection, IT security, and the willingness of citizens to use digital services across borders can only occur effectively when individuals are confident (trust) that their personal data will be effectively protected (Commission, 2017c).

This is particularly important as a combination of big data, analytics, and artificial intelligence supports the development of what could be termed 'decentralised autonomous services' and 'anticipatory policymaking' where automated systems, using new technologies such as blockchain, make decisions without human intervention. However, as a report for the European Parliament warns: "*the decentralised, cross-boundary character of blockchain raises jurisdictional issues as it seems to diffuse institutional accountability and legal responsibility in an unprecedented manner, rendering the need for a harmonised regulatory approach at the transnational level more pertinent compared with a local or regional one*" (Boucher, 2017). Figure 1 summarises the EU activities.

Figure 1: EU Big Data Landscape



Confidentiality and privacy are particularly important when children are involved. There are organisational challenges, such as how education systems can modernise sufficiently to have the analytical capabilities in place to cope with big data opportunities and challenges (Claros and Davies, 2016). Recent research for the Commission shows that across the EU substantial changes are taking place in digital learning, and over half of respondents to a survey noted that digital technologies and online learning had impacted on the curriculum and assessment mechanisms (Shapiro et al., 2016).

Nevertheless, to cope with the flow of new types of data from IT-enhanced learning educators will need to undergo a major 'cultural change' in their IT skills (OECD, 2016f), use of data, and their teaching methods, which will require significant levels of training and re-training across the teaching profession (Shacklock, 2016), and maybe even a transformation of what a teacher 'is', beyond being an officially accredited professional.

Institutions can feel threatened by disruptive developments in education. At the tertiary level, MOOCs have been viewed as potential curricular and the recognition of threats, rather than opportunities, although the recognition of MOOC learning by ECTS credits has been slowly growing (Whitthaus et al., 2016). To help in the recognition of learning outcomes and qualifications at tertiary levels the EU does have the benefit of NARIC[18] Centres that provide professional mediation between international students and HEIs in the area of the validation of qualifications.

The request from the Commission identified some important common issues across the big data landscape. The data sources are large, often require complex analysis, and are produced across a very wide range of producers (in itself a significant challenge since to date strong reliance has been put on data from a relatively limited number of official statistical providers) and forms (going far beyond censuses, surveys etc.). Linking existing data sources more effectively and efficiently requires mechanisms to link the data (to make the resulting services interoperable), and in March 2017, a new European Interoperability Framework was released, providing guidelines to MS to help "*ensure that their services are standardised, automated, streamlined and provided securely in less time and with less effort*" (Commission, 2017b)

The Commission has been active in promoting debate and discussion about the opportunities and challenges relating to big data. The Joint Research Centre has produced DigCompOrg, a European Framework for Digitally-Competent Educational Organisations (Commission, 2016b), to help education organisations adapt to new technologies and information use. The platform 'Big Data Europe[19]' is stimulating debate and discussion in the field of Climate, Energy, Food, Health, Transport, Security, and Social Sciences, although it does not yet provide a strong focus on education.

The Commission's Big Data Strategy[20] was introduced in July 2014. The Strategy was primarily aimed at the development of the Digital Single Market[21], focusing on challenges relating to cross-border information flows, cloud computing, security and trust (Commission, 2014d) to deliver "*accelerated innovation, productivity growth, and increased competitiveness in data across the whole economy, as well as on the global market with Europe as a key player*".

Big data has been a specific focus of calls for proposals under Horizon 2020[22], with some of the research themes being of relevance to this study: cross-sectorial and cross-lingual data integration and experimentation; large scale pilot actions in sectors best benefitting

[18] https://ec.europa.eu/education/resources/national-academic-recognition-centres_en
[19] https://www.big-data-europe.eu/
[20] https://ec.europa.eu/digital-single-market/en/towards-thriving-data-driven-economy
[21] http://ec.europa.eu/futurium/en/digital4eu/ideas/big%20data
[22] https://ec.europa.eu/digital-single-market/en/news/big-data-information-and-networking-days-horizon-2020-topics

from data-driven innovation, support, industrial skills, benchmarking and evaluation; and privacy-preserving big data technologies.

The potential use of big data across wide areas of policy and knowledge is being explored. Eurostat are considering the role of big data in the 'Statistical Office of the future' (ESTAT, 2015). Eurostat and the European Statistical System (ESS) are focusing on the potential for big data to improve official statistics. The Scheveningen Memorandum on "*Big Data and Official Statistics*" was adopted by the European Statistical System Committee (ESSC) on 27 September 2013, encouraging all partners to explore big data usage (ESTAT, 2013). One of the projects is ESSnet Big Data, which runs until May 2018 explores how big data can contribute to the regular production of official statistics (ESTAT, 2014, ESTAT, 2016b).

Commissioner Moedas has promoted big data as being of potential value in providing better forecasting, and more effective measurement of research impact on innovation (Moedas, 2016). In the context of education, Commissioner Navraciscs has highlighted the partnership of research and education in achieving better evidence-led policy (Navraciscs, 2016). The Commission has been energetic in facilitating dialogue across research, education, and innovation. For example, a public-private partnership was established in October 2014, bringing together businesses (ranging from large companies to SMEs), researchers and academics "*to cooperate in data-related research and innovation*". They will use four instruments within Horizon 2020: large-scale demonstrators; innovation spaces to bring developers and end-users together; technical projects to accelerate the use of key enabling technologies; and, networking and community building (Commission, 2017a).

The EU policy actions have been strongly addressing the themes in this study, and they contribute towards achieving more effective monitoring of education systems: a need to ensure data production is fully inclusive of all students, and that there is equity resulting from analytics; a need for effective skills forecasting that informs education systems of the needs of the labour market; assessment and student tracking mechanisms that ensure individualisation of teaching and learning; and an effective and responsive regulatory and ethics structure to ensure that privacy and security help to maintain confidence and trust in the use of highly detailed individual data.

Key points from the expert contributions

Students' privacy

Privacy and data protection provide fundamental frameworks for any big data developments in the monitoring of education systems. Any developments will need to consider the values and laws that enable societies to (continue to) function on the basis of fundamental rights and democracy. These values and laws include equity (the topic of another section), and privacy and data protection.

There are particular risks in the collection, processing, and sharing of personal data. Being monitored can produce 'chilling effects' where people feel obliged to conform, or a feeling that there is a 'big brother' form of surveillance, for example where CCTV cameras are used in classroom environments. For organisations, there may be a temptation for 'mission creep', and use data beyond their original monitoring purposes. Detailed individual data can allow for inferences towards health status, psychological variables, and other sensitive data, captured in fine-grained profiles that may lead to stigmatization, discrimination and exclusion. With big data there is the potential for educational data collection to extend into increasingly private spaces: while the classroom is already a non-public space, educational software on mobile phones may easily collect data in highly private settings such as the home.

In considering how the risks can be addressed, this contribution argues that it is important to: a) build on current understandings of the notions of privacy and related concepts in order to anticipate the risks that may persist and/or unfold over a 10-30 year horizon; b) not only look at the direct interplay between 'big data' and 'privacy', but also critically investigate some of the promises associated with big data in education; and c) focus on data protection law and its comprehensive consideration of risks to the rights and freedoms of individuals. The overarching focus is on the right to protect data, and the EU General Data Protection Regulation (GDPR) is the law designed to ensure this for three reasons.

First, persons including learners, teachers, administrators, and are key elements of every educational setting. In the case of primary and secondary education, parents are also often involved. Therefore, each big-data intervention can create personal data, and lead to the processing of personal data. Each collection and processing of personal data, in turn, may interfere with the rights and freedoms of individuals – the rights to data protection and privacy, but also others such as the right to non-discrimination or freedom of speech.

Second, it is all rights and freedoms (to the extent that they are affected by data processing) that the GPDR aim to protect. Privacy is regarded through the lens of data protection.

Third, the GDPR is a comprehensively thought-out model that provides a clear framework for action so that education system monitoring can be effectively developed in the coming decades, while providing robust privacy protection in a way that is (as far as possible) technology-neutral, and not reactive to technology developments.

The section then reviews the elements of the GDPR before considering big data issues. Data may relate to which documents a student accesses on an educational platform (for example Classroom Management Software - CMS), and what they produce (e.g. answers in tests, forum posts). There are analytical data such as grades, teacher feedback for students, or records of absences of students that resemble entries in traditional student files, but are now electronic, and can be stored in School Management Software (SMS). There also are analytical data collected for educational monitoring. Traditionally, this is not at the level of the individual, but at the level of a class, school, school district, etc. (such as: number of hours of specific teaching activities), or aggregated to such a level (such as: percentage of pupils with a migration background, or achievements in a test such as PISA).

The potential significant expansion in coverage and resolution enabled by a big data approaches does not imply that 'more data means better monitoring'. There are challenges relating to data quality, measurement systems, how data can be effectively anonymised, whether long-term coherent data for learners in life-long 'eportfolios' risks pervasive surveillance, what types of big data can be used for 'effective' monitoring, who owns the data about students, what are the competences of the education institutions and their staff in analysing and interpreting big data, and whether a reliance on automation and analytics means that education pathways may become determined by private actors that are not subject to democratic control.

The section strongly advises that data protection and privacy must be designed into the big data systems and processes. For every planned data collection, software or hardware deployment, etc., it is imperative to take recourse to a fundamental principle of European data protection law, that of data minimisation. Can the same (e.g. learning or monitoring) effect be achieved with the help of less data? This amounts to a necessity test for proportionality testing.

The GDPR includes the mandate for a data protection (or privacy) impact assessment (PIA). This involves the identification of key stakeholders and their interests in a

proposed new technology or method, as well as how their rights could be affected. Which information flows, from where to where and how? What are the roles of the stakeholders in providing, disclosing, collecting, using and sharing the information and the purposes and outcomes of analytics? Based on the state of the art in privacy-enhancing technologies and processes, how can the negative effects be mitigated?

This last step amounts to applying data protection by design, which is however not limited to deploying certain technologies, but also involves organisational measures, and which utilise current algorithmic and procedural developments for making processing discrimination-aware, transparent, and accountable. The section concludes with a detailed elaboration of how the EU policy frameworks and tools can be used to maximise privacy by design in the big data environment.

Looking ahead 10-30 years it is important to consider how the EU can further strengthen the checks and balances that are already in place, and how they can be developed to be better able to cope with the immediacy (time) and granular scale (individual) of data produced in particular by learning platforms.

Data collection and analysis may be governed by what is technically feasible (e.g. comprehensive sensing, real-time data analysis and purely algorithmic decision-making) and economically plausible (e.g. a missing, poor, or one-sided evidence base concerning efficacy, lack of data post-processing), and it may be motivated by short-term economic objectives (e.g. fitting learners to jobs) and means (e.g. handing over educational data collection, analysis and decisions to the private sector). In such scenarios, there is no time for a 'reflective' consideration of legal implications such as breach of privacy. Instead, privacy and data protection needs to be designed into the process from the outset.

Data protection by design is mandated by the GDPR. The GDPR can provide such a foundation, which can then provide more trust and credibility for big data analytics since those whose data are being processed could be more aware that EU-level checks and balances are designed into the systems.

This argument sets the foundation for the other four key areas analysed. It draws a 'line in the sand' to argue that the fundamentals of data protection and privacy must be designed into big data systems used in education systems in the EU. If the learning platforms and the analytics have data protection integrated into their design, then national education systems may be in a better position to use big data to deliver more equity and efficiency, through individualised monitoring of assessment, better tracking for grouping students, and for this to be undertaken life-long (skills).

Educational efficiency and equity

Efficiency in an education system concerns the relationship between inputs and outputs, involving the efficient allocation of resources and particularly, balance between different kinds of resources, and the efficient use of these resources, making the best use of each particular resource. Equity is the extent to which all students can benefit from education and training. Education systems are equitable if they first ensure that the outcomes of education and training are independent of socio-economic background and other factors that lead to educational disadvantage, and secondly, that treatment reflects individuals' specific learning needs.

Efficiency and equity involve balancing policy priorities and resources. At one end, there are finite financial resources, and achieving equity is not simply a matter of injecting uncontrolled amounts of funding into a system, since the way a system 'performs' (quality of teachers, quality of teaching pedagogy and content, learning technologies etc.), will influence the ways in which resources are applied. At the other end, there are the needs of individual learners that are often complex to identify (for example where the

family and health circumstances of a child change, requiring multiple agency inputs), and then for the needs to be resourced.

There is not a direct causal link between particular interventions to make things more efficient or equitable, whether it is the introduction of new teaching technologies or pedagogies, early childhood education and care, well-trained and innovative teachers, or strong school leadership,

Central in the comparative monitoring of equity and efficiency have been the testing programmes initiated by the OECD and by the International Association for the Evaluation of Educational Achievement (IEA). While the international testing programmes provide good country-level comparability, and they have significant political influence at country level, the assessments are time-based (relating to a particular year, and with results usually not published until the following year), and it can be difficult to link the learning lessons to the delivery of equity, in real time, and at the individual level of the student.

The end result of existing approaches to monitoring has been a time lag, where policy decisions and system evaluations will be based on older, and potentially out-of-date, data. Equity and efficiency are therefore critically important considerations for education systems, but have been difficult to monitor consistently (in a timely manner) and also comprehensively (monitoring in detail at learner level in a way). Policy makers need to know how their education system performs against others in the world, and to identify what policies and practices they can consider to help improve it. They also need to understand how their education system effectively delivers equity to all learners, irrespective of their needs. It requires data that are longitudinal, and capable of providing 'early warning signs', which provide feedback loops to parents, communities and policies. The section then reviews the predominant approaches to monitoring equity and efficiency, exploring emerging approaches using big data, and looking to the future potential for monitoring efficiency and equity in a holistic and individualised manner.

A range of issues need to be considered, including the harmonisation of monitoring data across education systems, the need for the 'right questions' to be asked when monitoring efficiency and equity, since these will determine the data to be collected – again the theme of 'more does not mean better' returns. Since equity requires a focus on individual student needs, there is a need for data and analysis to be sensitive to local and individual dynamics, while still needing to be capable of harmonisation and comparability when monitoring the education system. Efficiency is monitored at present through processes such as standardised school inspection (looking for example at absolute attainment standards or value-adding), performance management of teaching and management staff, and can also be addressed through organisational change and system modernisation.

The section focuses on examples of big data developments, ranging from the large-scale integration of data by the World Bank, in the UK, and then details how a huge range of data sources are shared and analysed through interoperability processes in Belgium. Two specific big data developments are presented, first for Portugal where the Troika required the development of a new education monitoring system. Second, a fully operational big data approach is shown for Estonia, where data are fully integrated from student level to system level, where students and parents can access learning progress (transparency of process and openness of data), and where (acknowledging the arguments in the following privacy section) underpinning trust and confidence in the system are the powerful security and privacy protections applied to the national identity card, with very strong system security and cyber-defences.

The section looks at how a big data approach could first (in the next 10 years) mandate 'privacy by design' (the privacy section develops this), and develop the interoperability frameworks, while supporting those Member States that wish to build on the systems

already in place in countries such as Portugal and Estonia. The widespread adoption of an 'atomic' level of student monitoring would not just look at their educational performance through learning platforms, but would also look at issues of value adding, personalised and blended learning, flag learning issues by relating them to data from other relevant data systems and assess whether the learning issues are purely educational, or are influenced by other social or external issues. It could change the nature of what is a 'school' as an institution, while overcoming the friction of location and distance by linking the equity needs of students to resources delivered through learning platforms.

Assessment

This section examines the potential of big data in the assessment process to enhance the quality and monitoring of education systems. It covers the ways that assessment is both undertaken and 'assessed' in new teaching and learning systems, noting that the international nature of many of the systems means that assessment information may be stored and processed beyond the 'borders' of national education systems.

Educational data now covers a much broader range of activities, such as on-line student social interactions, including text, audio, and video data, and fine-grained interactions, while solving authentic assessment problems. Over recent decades teaching and learning has moved student learning to a more cognitively active process, for example with rapid feedback, paths for remediation, and the ability for students to self-pace. Such pedagogies, have delivered significant learning gains. However, to date, delivery of such gains has required costly investment in human capital (teachers etc.) that is located in traditional education institutions.

A more cost-effective approach is through learning at scale: having substantial course resources (increasingly, entire courses) shared by thousands of students. If a 30% improvement in learning outcomes could be achieved, then in the USA high school students would graduate with knowledge bases equivalent to current college (university) graduates. In evaluations of the edX platform, results showed significant learning gain in on-campus use. In a blended learning trial at San Jose State University, course completion rose from 59% to 91%.

However, the new systems bring challenges, since students using online learning platforms share personal information, and exchange views that are controversial and which would not be recorded in conventional assessment systems. As a result, educational technology has moved the learning experience from a space where such data is relatively safe, to one where it contains highly private information which could be damaging to students' future careers, family lives, and psychological well-being.

Furthermore, the data being gathered in the new IT systems are mostly being gathered by for-profit corporations. Educational data are considered proprietary, and while the bill for developing such technology comes indirectly from taxpayers, few corporations share it with governments, researchers, or the students to whom such data pertain. Students and teachers merely have access to aggregate results. Since such data are divided among hundreds of educational technology corporations, there is no easy way to combine or correlate such data.

While sophisticated assessment is possible using learning platforms, as the data become aggregated (and this depends on an ability to merge data across many commercially proprietary platforms) the sheer volume of data challenges the researchers who may aim to analyse it at the education system level. Data from at-scale platforms is currently in the terabytes. Once multimedia data are included (for example, student conversations over video conferences), it will move into the petabyte or exabytes scale.

Furthermore, unlike the time-frozen data from the past, new big data have great velocity. The time-to-insight speed is essential. Education, as most fields, benefits from continuous improvement. Post-hoc analysis can drive policy choices. Semester-to-

semester feedback can help drive improved course design. Day-to-day feedback can help instructors identify where students are struggling, and provide feedback. Second-to-second feedback allows feedback to be provided just-in-time to students, helping them identify problem areas, and remedying knowledge gaps and misconceptions. Processing data in real-time at the velocity coming out of at-scale learning systems is still an area of early research with challenges difficult even for highly skilled computer scientists.

Monitoring at the education system level also requires a new understanding of the limitations of big data – big does not imply better, and nor is it easy to anonymise data sufficiently to avoid 'de-identification' and breaches of privacy. For example, once there are performance incentives (beyond simply using such data to inform teacher actions), there are incentives to game the big data systems. Students, especially more affluent ones, may take test preparation courses whose primary goal is to train students in test taking to bias their results, increasing socioeconomic advantages.

Looking ahead over the next 10-30 years, educational resource production, data collection, and educational technology is likely to become more unified and centralised, and blended learning will be the norm. Active learning activities and online assessments will result in superior outcomes in both student learning and engagement. The role of the teacher is shifting from the primary source of information to working with students 1:1 utilising such digital materials.

The production of blended learning resources will be driven by economies of scale. Economics drives curriculum, course, and educational resource design to be centralised. It's a natural monopoly, and there are fixed costs to creation, and near zero incremental costs to additional usage. It's not just a natural monopoly – it has strong network effects. A platform with more students and teachers has access to more data, to more contributions from teachers and students, and to a more diverse group of students. Student forums have more activity.

Governments spend huge amounts of money on their education systems, and this sector is seen as a significant business opportunity. Major corporations have education initiatives, and investors have financed the three major MOOC initiatives, Udacity, Coursera, and edX, at a level of over a third of a billion dollars, and their commercial valuation is many times that level indicating strong business confidence in making money from learning.

In many ways, the landscape of educational technology resembles that of computing circa 1975, or e-commerce circa 1999. There are many competing platforms, and while it is too early to tell which ones will dominate. But, without appropriate regulation, if a monopoly is in place, progress generally stops, and that is where there are policy challenges for the EU in the areas of data rights, privacy, security, research access and analytics, inclusion and equity, and (with increasing internationalisation of platforms) cross-border regulation.

Student tracking

Tracking in the context of this study is the process of grouping students by ability. The relationship between forms of tracking used to inform educational policy and practice is important, and micro-level behavioural observations can be used to transform practice in ways that achieve macro-level policy objectives. Ultimately, understanding student behaviour and problems with learning can feed back into policy decisions about where and when students should learn specific skills.

Thus, these innovations have the potential not only to improve the efficiency, speed and accuracy of policy forecasts, but also to transform the educational practices that underpin policy implementation. For example, groups of students with different abilities could be tracked and streamed into different educational trajectories based on predictions of future labour market's needs. A key to the success of this vision is the measurement and

use of appropriate big data.

Until recently different sorts of data were gathered and used for different purposes and held in separate databases so were difficult to combine. However, the potential of tracking systems can be seen from those countries that currently embrace them, such as Singapore. At the end of primary education students take an assessment which will decide the type of secondary school that they enter, and ultimately the kind of qualifications that are received. Multiple schools in Europe have also applied some variety of tracking. The Netherlands, for example, has an education system that utilises streaming from a young age and is ranked among the best of European educational systems according to international assessments (PISA). Similar to Singapore, an aptitude test is administered at the end of primary education to guide teachers and parents in recommending what type of secondary education to pursue.

However, tracking is controversial. Research has found that ability grouping does have a small positive effect on student achievement, but that it is one of the least effective approaches to increasing student capabilities, that is some cases the inequality gap in terms of student achievement was consistently worsened in educational systems using tracking. Tracking risks the negatively impacting the majority of the student body, due to polarisation of top and bottom students, high achieving students being forced to advance at a rate that is too quick, and a lack of pedagogical variety due to perceived homogenous classrooms.

The use of big data to inform algorithms, that would suggest paths for students, could tackle some of the biggest challenges currently faced by tracking. However, the effectiveness at creating educational equality and actually improving results should be carefully monitored. The section reviews some of the challenges associated with the sourcing, storing and analysis of big data. Data sources often are distributed across multiple sources and servers, and are gathered using different methods. To enable comparability, data must have standardised quantitative and qualitative indicators that offer insights policy and practice.

Gathering micro-level student data is complicated, not only because the variables are complex, but also because it requires intensive data-gathering and real-time analysis, and the analytic algorithms used will strongly determine what data are used and how they are processed. The available data is mostly limited to online learner activity, and does not offer direct evidence of offline activity and cognitive development. Data are often held in 'silos' at difference stages of lifelong learning.

There are important legal and ethical implications associated with the use of student data such as transparency, consent, and rights to seek redress. For example, it could be beneficial to use social big data, utilised by a tracking system that is integrated with social media. That has the potential to help individuals and students in ways outside of the classroom. Teachers and schools do not have the time to follow each of their students using social media, but big data based tracking would make it possible to have recommendations about which students are potentially the targets of cyberbullying. This could be taken into consideration in addition to academic achievement when recommending students for different tracks.

Therefore, a major challenge for education systems in Europe is to implement student tracking in ways that enable us to extract meaning from large datasets being generated through micro-level, online student activity and to distil this data into usable information for students, teachers, and governments. The policy questions at the EU level will need to address whether student tracking will promote equity, and whether 'success' can be more richly assessed with a wider range of data.

Policy will need to understand whether algorithms are objective and inclusive in their tracking of students, or whether new social and educational divides are being created.

Research will need to assess the efficacy and effectiveness of using data from social media, and who 'owns' such data, as well as data in the learning platforms. Data privacy and data protection (covered in the privacy section of this study) should be based firmly in ethical guidelines and practices.

Looking ahead 10-30 years the role of big data in tracking students will depend on how tracking is being used. It will need to adapt to rapid changes in the ways students learn, for example tracking how they learn and linking them to other students who learn in similar ways. Governments will need to rethink their educational models, and the role that tracking can play in them as fundamental changes in education systems will likely be needed. If policies are enforced that underpin education with systems based on conventional educational models, an opportunity will be missed for future development and change. Before investing funding into tracking, governments must consider fundamental questions about the effectiveness of tracking and where tracking will fit within evolving educational practice across the EU.

Skills forecasting

This contribution first identifies ways in which big data can be implemented in the analysis of labour market demands. Second, it outlines possible avenues of using educational big data in helping develop students' skills and to improve the responsiveness of educational systems to labour market skills demand. Finally, the opportunities and challenges are taken into possible actions at EU level.

Better coordination between education and the labour market is needed to overcome major skills challenges, such as skill shortages, mismatches between job requirements, and the education levels and skills of those in the jobs. Big data and analytics have the potential to help skills development across educational pathways which can help increase pupils' and students' academic performances and help them make personalised career choices, matching better their education outcomes, skills, competencies, and the needs of the labour market.

However, developments to date in linking skills, labour market, and education are limited. Big data are used in job vacancies and recruitments, for example through social media, and in education learning and school management systems are building rich understandings of the educational progression at the individual level. There are opportunities with big data for better monitoring of the skill needs of labour markets and linking it to more responsive and data-rich monitoring of education systems.

In building a better understanding of skills forecasting big data is faced with particular challenges. There are questions about representativeness, whether the vacancies posted on online job portals are fully representative of the labour market, and whether they are communicating an accurate picture of needs to education systems. In addition, the descriptions of jobs and skill requirements are not standardised, nor are the big data relating to students in the education system concerning their skills and competencies. IT infrastructures and data analysis capacity across education is still a challenge, at least in a short to medium term perspective. The contribution notes some individual projects and initiatives that are using big data to build better links between education, skills and the labour market, it further notes developments in big data and technology platforms to overcome data fragmentation.

Recommendations for policy action focuses on the creation of a robust infrastructure and methodological framework enabling data collection and data analysis, and regulatory frameworks for privacy and governance of big data for education and skills. This can be undertaken through facilitating collaboration between labour market bodies and social partners, encouraging the digitisation of the job vacancies which involve lower qualifications and vocational skills, integrating big data analysis as part of the Cedefop and Eurostat work on skills forecasting.

Studies and pilot projects could provide examples of what can be achieved, and can also be used to support capacity building, implement big data into education curricula in a way that supports building consistent knowledge about skills, exchange good practice, and undertake the monitoring of emerging big data sources. Collaboration would also be valuable between education ministries, Eurostat, and national statistics offices to build coherence and harmonisation across big data.

The contribution finally looks ahead to future scenarios, noting the potential for significant diversification and fragmentation of education structures and the need for governments to work on supporting mechanisms to build big data across fragmented structures.

Looking ahead

In looking ahead, the study was asked to consider:

> *"how advances in big data are likely to transform the context and methodology of monitoring educational systems within a long-term perspective (10-30 years) and impact the evidence based policy development in the sector"*.

Each contribution takes its own perspective on the future. Looking forward, the latest (2016) Gartner Hype Cycle in education seems pessimistic about big data, seeing it among other aspects such as adaptive learning platforms, open micro-credentials, and gamification as "*Sliding Into the Trough*" (GARTNER, 2016) after initial high expectations. The activities that are more successful in the hype cycle tend to be digital applications of conventional educational processes such as digital assessment, learning analytics and competency-based education platforms.

Technology development can be easier to predict than technology adoption. Big data utilisation will need to be part of a process that does more than computerise what exists. There will be fundamental changes in the relationships between teachers, students, and families, and all will need to be part of the learning process. Curricula will need to break across subject boundaries, and assessment will need to be less subject focused, and to "*cross the borders between subjects, between 'academic' and 'vocational' learning, and between the worlds of adults and students*" (Hampson et al., 2016).

Looking forward also involves looking back over the past decades, to see how past predictions foresaw challenges and opportunities of the present, particularly in the context of technologies and education policies. A Commission resolution back in 1976 had clearly identified that education systems needed to have a much stronger link on "*the transition from school to working life*", and that it should be inclusive. Too many young people were not well equipped with knowledge, skills, and competences needed for the labour (Commission, 2006b).

The development of an information and knowledge society was also well-established in the 1980s, and 1990s, and the Commission noted that there would be significant changes in the balance of work and leisure, and that there would be the potential for quality improvements in education systems (Commission, 1994). The significant changes in the balance of work and leisure noted in 1994 had been acknowledged in a 1993 White Paper on "*Growth, competitiveness and employment: the challenges and ways forward into the 21st century*".

Here, the Commission was clear that a major challenge would be the reform of education and training systems, with a need to focus on lifelong learning, with continuous learning opportunities from basic education to training that was available throughout working life (Commission, 1993). A goal clearly was to join up the distributed components of education systems in a way that allowed an individual to 'carry' their education and training outcomes with them throughout their life – robust, detailed, and comparable data would be important in monitoring the pathway to achieving the goal.

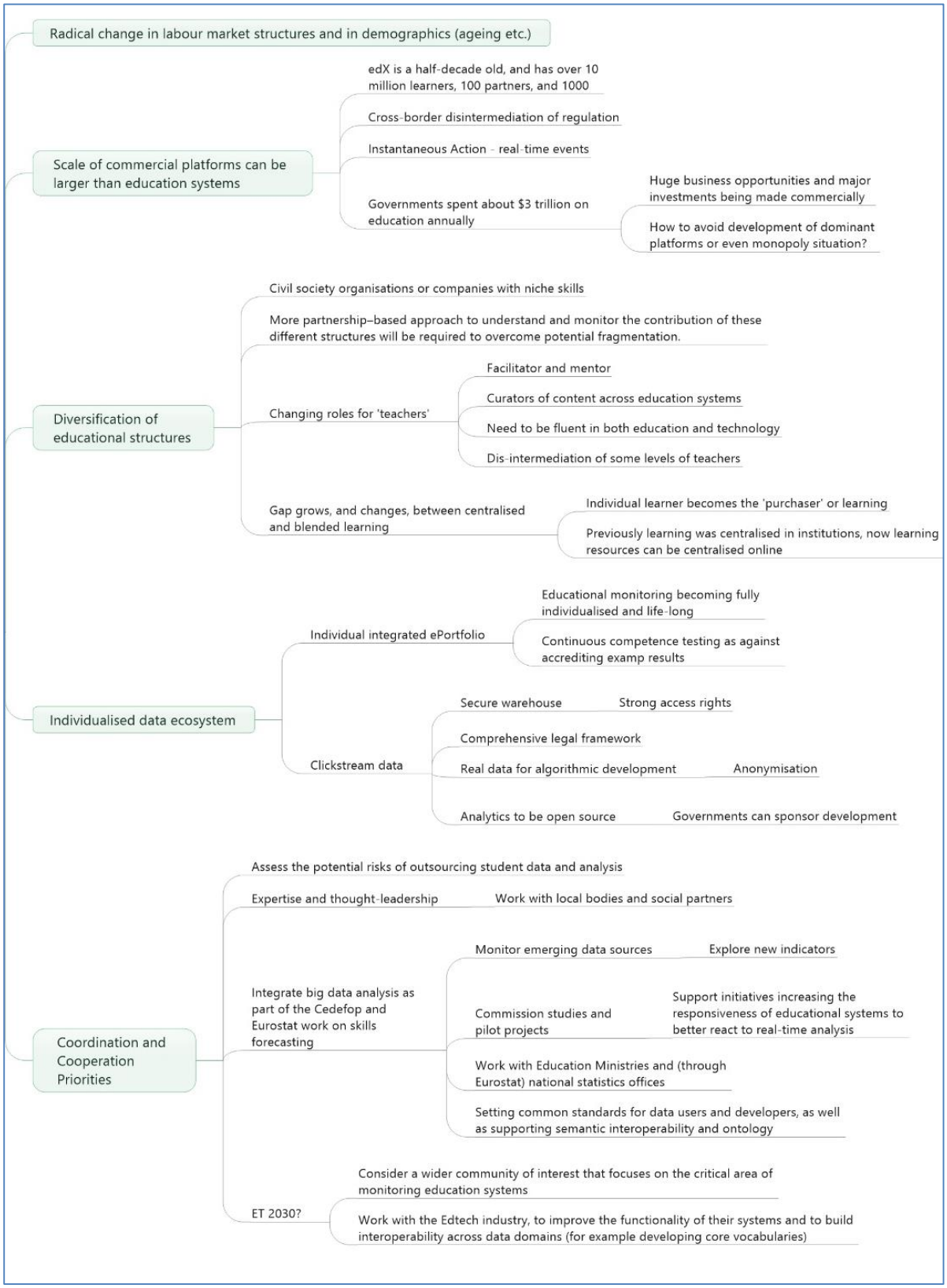
The EU had also been putting resources into building robust and harmonised statistics, with Eurostat starting to publish education statistics in 1978, subsequently working with UNESCO and the OECD on classification systems. Particular impetus was provided through "*the adoption on 5 December 1994 of a Council resolution on the promotion of education statistics in the EU*" (Commission, 2006b).

What was further foreseen by policy makers was increasing access to information at an individual level, and that the information would no longer be filtered by 'authority figures' such as schools, public libraries, newspapers, or the mass media. Former US Vice-President Al Gore talked about the 'information superhighway' in the 1990s, and in the context of learning considered a vision where a schoolchild could return home, switch on a computer, and access the entire Library of Congress (Gore, 1994).

This was looking ahead to a world of increasing self-learning, of access to learning resources for anyone connected to the internet. MOOCs, OER and connected classrooms are logical outcomes of the information superhighway. However, it was a vision that in many sectors was then accompanied by major disruptive change, including organisational change in government services, commerce, and globalised companies which governments have struggled to cope with in terms of regulation.

In 2017, we are in an environment (Figure 2) of educational mobility, education and training transparency tools, eCommerce, eGovernment, remote working, wireless internet, robotics, social networks, the cloud, civilian access to GPS location, and disruptive business models such as Uber and Airbnb. These changes have also been associated with challenges such as the automation of many jobs, digital divides, changing employer/employee relationships (such as casualisation or zero-hours contracts), regulatory risks (where new business models threaten existing regulated business such as taxis and hotels), business strategy (globalisation and extended supply chains), and risks to individuals from their data and privacy being threatened by cybercrime and IT security failures (Simonite, 2016, Baraniuk, 2016).

Figure 2: Big Data Developments for Education Systems



Yet, while radical disruption has been evident in many sectors, radical organisational change has not been as evident in education systems. The focus of compulsory education at primary and secondary levels (Eurydice, 2016a), remains strongly oriented to formally structured school years (Eurydice, 2016c) based on physical presence at school (Eurydice, 2016b), with nationally or federally specified curricula (Eurydice, 2016d), and on the same age-based transitions from primary to secondary to tertiary education.

While there is often connectivity of curriculum from primary to secondary, this tends not to be the case for secondary to tertiary, and life-long learning also has little connectivity. So, one feature of education systems in the age of big data is that while they have 'modernised' their teaching and learning within education layers, the borders between the layers are still very age-related, rather than lifelong learning oriented. That is not a particular concern if the **data discontinuities that exist across the borders of education levels** can be overcome, and big data shows considerable potential.

Looking ahead, it is unlikely that existing information structures in education system structures will remain fit for purpose, as lifelong learning and continuous education and training through a long working life become more common. Education systems will therefore need to be radically restructured, and much more informed by robust, detailed, and timely evidence, and big data will play a vital role.

There could be two potentially converging approaches to fully modernising education systems. First, **restructuring and organisational change** can continue to happen, although this will need to engage with actors such as teacher unions as teaching and learning patterns change. Paradoxically, the education systems which are experiencing the most challenges over quality and labour market relevance could be the ones most suitable for change – their legacy effects will be less than those more advanced, but still conventionally structured, systems.

Second, **big data integration and transparency tools could enable learners to 'transport' their learning across all the layers of education, from primary to life-long learning.** For example, from formal to informal, and from education that leads to a formal qualification, to a collection of recognised digital badges. The latter process would be strongly reliant on the use of big data, ePortfolios, and more flexible systems of recognition (supported by robust classification systems for skills, competencies, qualifications etc.). To achieve this big data analytics and developments in artificial intelligence will be important.

Students' privacy

Big data has the potential to change education and education monitoring in wide-ranging ways. This study investigates the promises and risks of these changes with regard to student tracking and other data collection and processing for purposes of student assessment, grouping, and skills forecasting. However, whatever uses of big data technologies European decision-makers envisage: they need to consider the values and laws that enable our societies to (continue to) function on the basis of fundamental rights and democracy. These values and laws include equity (the topic of another section), and (important when talking about big data) privacy and data protection. This material investigates possible futures for the use of big data educational technologies and the privacy of students, in the context of monitoring education systems at European, as well as more local levels, down to individual schools. The section concludes with recommendations for policy, in particular stressing the need for data protection by design as a fundamental pre-requisite for any education big data systems.

Big data, educational technology, and opportunities and threats in an EU context

The Commission's definition of big data "*refers to large amounts of different types of data produced with high velocity from a high number of various types of sources*" (Commission, 2014d). The theme of student privacy in the context of big data is fundamentally an issue of protecting and empowering students in situations where schools, education platform providers and education administrations collect, hold, and share extensive and highly integrated information about students. The promise is that, processed using powerful statistical analytics, the resulting large volumes of data can be used positively to improve teaching, learning and education monitoring, and to safeguard against abusive behaviour both by students and teachers. These expected advantages will be described in detail in other sections of this report.

Integrating massive amounts of information has been at the heart of many EU initiatives for providing better services for citizens, ranging from education to transport, from health to governance.

However, beyond supporting innovation and technology development, the Commission has a clear understanding that with greater integration of information, and the greater online availability of integrated information, come risks, particularly in the areas of technological security (and threats to weaknesses in security through hacking and cybercrime), misuse of information and communication (such as fraudulent use of information, or the unethical use of digital communication tools such as social networks and email), a lack of knowledge sharing among organisations and administrations, and threats to data protection and privacy.

The Commission strongly supports the development of privacy enhancing technologies (PETs) (Commission, 2007). It has been developing the security and trust underpinnings for integrated and interoperable services through the Connecting Europe Facility[23] (CEF) through the provision of 'building blocks' for European-wide electronic identity and signatures, invoicing, eDelivery of documents between public administrations, and automated translation.[24] It promotes the development of a Data Driven Economy, involving "*EU action to provide the right framework conditions for a single market for big data and cloud computing*" (Commission, 2014d).

The particular vulnerability and needs of minors are recognised through laws (see following sections for a brief overview) and other means. The Commission has developed

[23] https://ec.europa.eu/digital-single-market/en/connecting-europe-facility
[24] https://ec.europa.eu/cefdigital/wiki/display/CEFDIGITAL/CEF+Digital+Home

a cybersecurity cooperation platform[25], a Safer Internet[26] resource aimed particularly at children, and there are EU agencies such as the EU Agency for Network and Information Security (ENISA[27]), or the EUROPOL European Cybercrime Centre (EC3[28]). There is a single European phone number, 116 111[29], for children to call if they need advice or help.

CEF, data protection legislation, and the other Commission and related EU developments noted above, in effect represent a metaphor for a discussion about the use of big data within education systems. The risks associated with the promises of big data must be clearly understood and carefully confronted through combinations of regulation, electronic security, and sharing intelligence. The next section provides an overview of frequently considered risks.

Frequently discussed risks of the collection, processing, and sharing of personal data

The collection, processing, and sharing of personal data can be considered problematic, in general and particularly in educational settings. Some effects have been described in a growing literature, and three effects in particular are the most often described. Being observed can silence people and make them fall in line ('**chilling effects**', for an overview, see (Penney, 2016)), which is incompatible with democratic participation (Bundesverfassungsgericht, 1983) – a mindset that schools are supposed to help create.[30]

Data creates desires and invites **mission creep**, which can lead to further pressures and chilling effects. For example, Big Brother Watch considers that teachers must use student monitoring to enhance education experiences, but the software (or instructions once the software is in place) may allow or require them to "*spend their lessons monitoring student's computer screens for signs of inappropriate behaviour*" (BBW, 2016). Another example are wall-mounted cameras that can use facial recognition to see if learners are 'bored' and change the teaching directly (Kuchler, 2017), but the same cameras could be used to look at other facial gestures, or even to monitor gait or body temperature. Other examples may be welcome uses of technology for some, and mission creep for others, as when camera and data surveillance of educators and students is used to provide evidence trails in cases of accusations of abuse or misbehaviour (GAO, 2014).

In general, observations create rich datasets that, often together with other data about persons, allow for inferences towards health status, psychological variables, and other sensitive data, captured in fine-grained **profiles** that may lead to **stigmatization, discrimination and exclusion**.

The potential for widespread integration of student data leads to well-founded concerns over a 'big brother' form of **surveillance**. Already there have been hostile reactions to new learning platforms, such as intrusive monitoring of "*social habits, student attention span, and more*" (McIntyre, 2016). The US Electronic Frontier Foundation took concerns to the national level in 2015 when it complained to the US Federal Trade Commission that Google was breaching its assurances about student privacy by using the cloud to store information about what is now its "*G Suite for Education"* (Okuda, 2016). Further issues arise due to the widespread use of mobile devices. Via these, educational data collection extends into increasingly **private spaces**: while the classroom is already a non-public space, educational software on mobile phones may easily collect data in highly private settings such as the home (Warrell, 2015).

[25] https://ec.europa.eu/cefdigital/wiki/display/CEFDIGITAL/Cybersecurity
[26] https://ec.europa.eu/cefdigital/wiki/display/CEFDIGITAL/Safer+Internet
[27] https://www.enisa.europa.eu/
[28] https://www.europol.europa.eu/about-europol/european-cybercrime-centre-ec3
[29] https://ec.europa.eu/digital-single-market/en/116-helplines
[30] Cf. the preamble of school laws in this document

While this section focuses on the main challenges of protecting the privacy of students, and where educational environments increasingly have highly integrated and rich information resources about students (and also about teachers, management and educational organisations), there also are associated issues related to these stakeholders' roles in respecting the privacy and data rights of other students, teachers, managers, parents, etc.

Concrete cases may involve threats and opportunities. In 2003 the establishment of CCTV cameras in all classrooms in a school was controversial, with teaching unions fearing **pervasive monitoring of teachers**, although (in the context of comments earlier about safeguarding) "*they recognise the cameras may protect teachers from false allegations by pupils*" (BBC, 2003). The government in South Korea recommended that when a mobile phone is being used to take a photo it should make a loud enough sound so that the subjects would be aware that a photo was being taken (BBC, 2004a). The development of high-definition camera phones led to students being banned from taking them to school for fear of unethical use (BBC, 2004b). In September 2003, a school district in Pennsylvania (USA) started to fingerprint every student and use biometric identification (Graziano, 2003).

Privacy and data protection principles and laws aim at avoiding such effects by placing limits on collection, processing, storage, and sharing.

Approach

A discussion of big data, education monitoring, and privacy can take different approaches. This section will argue why it is important to: a) build on current understandings of the notions of privacy and related concepts in order to anticipate the risks that may persist and/or unfold over a 10-30 year horizon; b) not only look at the direct interplay between 'big data' and 'privacy', but also critically investigate some of the promises associated with big data in education; and c) focus on data protection law and its comprehensive consideration of risks to the rights and freedoms of individuals.

Dangers to privacy arise from whatever are – in the EU at least – possible or actual violations of **fundamental rights**. Specifically, this is about the rights to privacy (Article 8 of the European Convention on Human Rights, Article 7 of the Charter of Fundamental Rights of the European Union) and data protection (Article 8 of the Charter of Fundamental Rights of the European Union).

Fundamental rights codify ethical principles, and these change only slowly. The Universal Declaration of Human Rights dates from 1948, the European Convention on Human Rights came into effect in 1953 (and the Charter in 2000), and recent proposals such as the European Digital Charter[31] refer to legal developments such as the fundamental rights to informational self-determination and to the guarantee of the confidentiality and integrity of information technology systems, both formulated by the German Constitutional Court in the last decennia of the 20th century on the basis of the German constitution of 1949. While it is possible that European case law will continue to modify the interpretation of these fundamental rights, these changes are far slower than some technological changes.

Should, and may, harm arising from violations of these fundamental rights be **balanced** against benefits to be had from new measures, in our case big data in educational technology?

EU data protection law aims at such a balancing act. While the General Data Protection Regulation (GDPR), the new EU-wide law to come into effect in May 2018 (Council,

[31] https://digitalcharta.eu/sprachen/

2016b), begins with a clear commitment to the protection of fundamental rights (Recitals 1ff.), it also aims at:

> "*{removing} the obstacles to flow personal data within the Union*" (Recital 10), with the goal of "*{contributing} to the accomplishment of an area of freedom, security and justice and of an economic union, to economic and social progress, to the strengthening and the convergence of the economies within the internal market*" (Recital 2). It also recalls that the "*right to the protection of personal data is not an absolute right; it must be considered in relation to its function in society and be balanced against other fundamental rights, in accordance with the principle of proportionality*" (Recital 4).

However, balancing according to the principle of proportionality is more than a simple weighing of risks against promises. Instead, any interferences with a fundamental right by a public authority must be *prescribed by law*, pursue a *legitimate aim*, be *suitable* to achieve this aim, *necessary* to achieve this aim: that is to be the least intrusive means, and it must be *reasonable*, i.e. pass a proportionality test 'stricto sensu', which consists of a weighing of interests whereby the consequences on fundamental rights are assessed against the objectives pursued (balance of interests).

Just as it is important to be aware of the "*false trade-off between privacy and {national} security*" (Solove, 2011), we need to beware of false trade-offs between privacy and presumed nirvanas of technologies and big data inferences – in education as in other areas. Rather, it is necessary to start from a clear, legitimate, and legally prescribed aim and then ask whether some data-collecting and processing activities are suitable to achieve this aim.

Suitability is an empirical question, and evidence is needed. This is the central topic of the sub-section reviewing big data and EU data protection developments, and decision makers need to carefully inspect the evidence base of any educational technology's claims. In this section, further questions are asked pertaining to how data protection concerns themselves may interact with the quality of such evidence bases.

It is important to ask whether the means are necessary. Can the educational/education-related aim also be achieved with less intrusive means, for example with less data (the principle of data minimisation)? Whose and which interests are at stake, and why and can they be weighed against one another? Finally, and going beyond the law, no specific form of educational monitoring, educational technology, or big-data developments are "*inevitable*" or "*alternative-free*"[32] (Bigge, 2006).

The main focus here is on the **right to data protection**, and the GDPR as the law designed to ensure this. The reasons are threefold. First, persons including learners, teachers, administrators, and are key elements of every educational setting. In the case of primary and secondary education, parents are also often involved. Therefore, each big-data intervention into educational processes, as a rule, creates personal data and leads to the processing of personal data. Each collection and processing of personal data, in turn, may interfere with the rights and freedoms of individuals – the rights to data protection and privacy, but also others such as the right to non-discrimination or freedom of speech.

Second, it is all these rights and freedoms (to the extent that they are affected by data processing) that the GPDR aim to protect. The GDPR is thus focussed on, but not limited to, the fundamental right to data protection. Taking this perspective and keeping with the present report's focus on big *data* in education, privacy is regarded through the lens of data protection.

Third, the GDPR is a comprehensively thought-out model that provides a clear framework for action so that education system monitoring can be effectively developed in the coming decades[33], while providing robust privacy protection in a way that is (as far as possible) technology-neutral, and not reactive to technology developments.

Data protection and privacy in the EU: basics and key terms

The right to privacy states that everyone has the "*right to respect for his private and family life, his home and his correspondence*" (Article 8 of the European Convention on Human Rights, Article 7 of the Charter of Fundamental Rights of the European Union) and that public authorities must not interfere with this right except in the interest of "*national security, public safety or the economic well-being of the country, for the prevention of disorder or crime, for the protection of health or morals, or for the protection of the rights and freedoms of others*" (Article 8, Convention). Privacy thus involves a protected sphere in which individuals can freely construct their identities.

Mass surveillance is understood as violating this right, and only targeted interception of traffic and location data in order to combat serious crime, including terrorism, is justified, according to a decision by the European Court of Justice.[34] Violations of the right to privacy may, but do not need to, arise from the processing of personal data.[35] The right to data protection states that the processing of personal data must rest on consent or on a(nother) legal basis, must be for specified purposes, and that the individual has rights of access and rectification. Thus, violations of the right to data protection may, but do not need to, involve violations of privacy: for further discussion see (Gutwirth and De Hert, 2006) (Kokott and Sobotta, 2013).

The GDPR recognises that unrestricted collection, processing and sharing may severely affect people's private lives and place unreasonable constraints on the development of their personality, and thus violate their privacy. It also recognises that other rights such as freedom of speech may be affected. To prevent such adverse effects, the law is based on *principles*, and it affords individuals *rights* and places *obligations* on data controllers. The rights embody informational self-determination, foreshadowed by Westin's short summary: "*the right of the individual to decide what information about himself should be communicated to others and under what circumstances*" (Westin, 1970).

Core data protection **principles** enshrined in European law are fairness, lawfulness and transparency; purpose limitation; data minimisation; data quality; and security, integrity and confidentiality. In addition, *individual empowerment* is regarded as a central goal of European data protection law (Article29, 2010).

Under the GDPR, individuals have the following **rights** (see the summary in (ICO, 2017)):

- The right to be informed;
- The right of access;
- The right to rectification;
- The right to erasure;
- The right to restrict processing;

[33] The GDPR's predecessor, the 1995 European Data Protection Directive, will, in 2018 when the Regulation comes into effect, have regulated 23 years.

[34] http://curia.europa.eu/juris/document/document.jsf?text=&docid=186492&doclang=en

[35] "any information relating to an identified or identifiable natural person ('data subject'); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person" (Article 4 (1) GDPR)

- The right to data portability;
- The right to object;
- Rights in relation to automated decision making and profiling.

The data controller's obligations are mirrors of the rights (e.g. to inform, grant access), technical enablers (data security), procedural and organisational rules to safeguard these rights (e.g. privacy impact assessment, data protection officer, data protection by design), and the meta-obligation to be accountable for fulfilling the other obligations, i.e. responsible and able to demonstrate compliance, and to notify of data breaches. Further obligations attempt at relieving individuals of (some of) the burden of responsibility of exercising their rights (data protection by default).

These obligations bind public as well as private actors.[36] In the education sector, schools and school authorities are examples of the first kind; any third parties involved (e.g. companies offering learning analytics software as a service) are examples of the second kind.

Young people[37] are particularly vulnerable, and are therefore afforded specific protection. The UN Convention on the Rights of the Child prohibits any arbitrary or unlawful interference with the child's privacy (UN, 1989). The GDPR[38] (Council, 2016b) specifies that particular attention must be given to protecting personal data relating to children who are less likely to be fully aware of their rights, and of the potential consequences of the outcomes of the use of their personal data. The Regulation notes the need for parental consent, provides strong protection where data is used for marketing or to generate profiles, but also understands that parental consent should not be required where there are "*preventive or counselling services offered directly to a child*." The Regulation places a strong obligation on national data protection authorities to undertake robust promotional activities to raise public awareness and knowledge about "*risks, rules, safeguards and rights in relation to processing*", particularly addressing the needs of children.

In addition to the general risks listed here, which have been described in much detail elsewhere, there are also risks of big-data collection and processing that are more specific to school and education monitoring contexts. They are linked to both big data as a concept and practice, and to the exercise of the rights listed above.

Types of big data in educational contexts

The following subsections investigate a number of specific challenges to data protection viewed in this larger sense that do, or may arise from increased uses of big data technologies in education. It will refer to various forms of (big) data collection and processing as education-related activities. Big data may arise from different sources and purposes, and this may give rise to different privacy-related concerns.

The first are the operational data of a pedagogical situation, such as which documents a student accesses on an educational platform, and what they produce (e.g. answers in tests, forum posts). For example, Classroom Management Software (CMS[39]) platforms aim at providing fully digitally connected learning environments. CMS has strong selling points: "*see everything your students see and keep them on-task*".[40] A CMS has

[36] Differences exist and are specified in the law. For the purposes of the present discussion, we can abstract from these differences.
[37] The UN Convention on the Rights of the Child defines a child as a person under 18. The GDPR follows this definition, but adds lower age limits for specific protections (such as 16 for parental consent). A number of specifics and exceptions apply, but are not relevant for the present discussion.
[38] See in particular Recitals 38, 58, 65, 71 and 75, and Articles 6, 8, 12, 40 en 57.
[39] For an indicative range of offerings see http://www.capterra.com/school-administration-software/
[40] http://www.netop.com/edu.htm

functionality that ranges from being able to monitor the screen of each student, check for student activity (for example noting whether they are active, or checking for inappropriate activity), filtering web access, launching apps individually.

A CMS integrates information for the teachers and institutions. Activity is time-stamped and easily audited, and it is an electronic audit trail that can be retained by the school and analysed easily – much different from physical archives of exercise books, or written assignments. Beyond web-style access logfiles, various technology is being trialled, such as body cameras (Burns, 2016c), sensors capturing facial expressions, heart rate, posture, and pupil dilation (Herold, 2016), or apps installed on students' mobile phones which tracks how long they spend working, socialising, exercising and sleeping (Warrell, 2015).

The data may be internal or external to educational technology. An example of external data is when students blog on Twitter as part of their assignment, or post on Facebook (without specific assignment) about their learning. Such data is generally used for analytical purposes at a small data scale: teachers use them to grade students, and/or learning analytics are applied to them. Further examples of such data are lesson plans and mixed pedagogical/administrative school data.

These datasets are big data in two ways. Firstly, their volume and depth is far larger than traditionally. For example, log files give finer-grained data of activities than traditional human observation, and sensors in the classroom observe behaviour even more closely. Secondly, data may be stored and used beyond the original learning situation.

The second type of data is the analytical data such as grades, teacher feedback for students, or records of absences of students that resemble entries in traditional student files, but are now electronic. School Management Software (SMS)[41] moves beyond a 'classroom' environment to a whole-school environment (and beyond where the same software is used across learning institutions in an administration. Facilities may include (the list is taken from the product Alma[42]):

- Integrated calendars, resource assignment, reporting across teachers, students, cohorts;
- Financial monitoring and management;
- Scheduling facilities: timetable, transport;
- Attendance monitoring, alerts to teachers, carers, parents;
- Pupil biographical and health data (allergies, immunisations, medical conditions, behaviours) and contact details of parents, carers, doctors etc.;
- Course management facilities: Assignment management (e.g. ensuring that students are not overloaded with work across their courses), assessment tracking, feedback to students;
- Integrated communications: bulletin boards, emails, secure staff communication, text alerts, emergency communication;
- Reporting and analytics: data visualisation, customised reporting, at all levels from students to the institution.

The Estonian big data education platform (see the equity and efficiency section) is an example of this type of technology already deployed in an EU country. Compared to traditional student files, these are big data mainly by virtue of longer retention times (e.g., ePortfolios), aggregation and access beyond traditional units such as the school

[41] e.g. the US product http://www.getalma.com/, or the Belgian product http://www.smartschool.be
[42] http://www.getalma.com/features.html

(e.g., the UK National Pupil Database), and also by a higher degree of datafication than in traditional settings (cf. teacher feedback in eSchool, which is written and always accessible to parents, as opposed to communicated orally in personal meetings).

The third type of data is analytical data, collected for educational monitoring. Traditionally, these are not at the level of the individual, but at the level of a class, school, school district, etc. (such as: number of hours of specific teaching activities), or aggregated to such a level (such as: percentage of pupils with a migration background, or achievements in a test such as PISA).[43]

A "big data" phenomenon well-known from other domains is also observable here: the re-purposing of data for new uses, and/or the lack of a clear purpose from the outset. For example, operational data may turn into national archives, or data from the Estonian population register, which forms a context of the platform, are never deleted and are archived forever (Björklund, 2016, p.924). Already today, data on early leaving from education and training are, in some European countries, not only collected for statistical purposes, but to track and follow up on individual students (Eurydice, 2017a, p.8). Conversely, the increasing availability of the first two data types discussed above may suggest their (re-)use for education monitoring – giving rise to more statistical data with no extra collection cost and effort. Such openness and re-uses are often hailed as *the* potential of big data by its proponents. However, they are also at odds with classical data collection and data protection principles, and pose new threats.

In all cases, data may contain personal data about students, teachers, parents, school administrators, and anyone else involved in the process.

These types of big data give rise to different concerns, and it is important to move beyond the risks which dominate the public debate noted above. These concerns arise mostly with respect to data of the first type, and therefore tend to neglect the consequences of big data of second and third types.

The following is structured by whether risks concern "scientific" questions, institutional issues, or socio-political questions. Many of them cut across all types of big data. Through the re-purposing of data, the risks concerning one type will also affect other types. This material is then followed by policy recommendations structured by decision-making level.

Key issues and Challenges

The evidence base of big-data educational technologies in interaction with data protection

Data quality

Educational monitoring in general, and big-data analyses and inferences in particular, require high-quality data. However, **data quality** faces a number of challenges. The first set of challenges is not specific to education: more is not necessarily better, and even without any constraints on data collection, it is impossible to collect 'all' data. All data necessarily are measurements contingent on choices, technology, and ultimately interpretation. While this statement is extremely general, it is repeated here because its importance cannot be over-emphasised, and critical discussions of measurement and sampling issues are often lacking (Kitchin, 2014).

[43] For examples of indicators based on such data, see EURYDICE. 2017a. *Structural Indicators for Monitoring Education and Training Systems in Europe 2016 – Thematic Overviews*. European Commission. Published February 16. Available: https://webgate.ec.europa.eu/fpfis/mwikis/eurydice/index.php/Publications:Structural_Indicators_for_Monitoring_Education_and_Training_Systems_in_Europe_2016_%E2%80%93_Thematic_Overviews. [Accessed February 16 2017]. ibid.

One risk of fine-grained measurement is the **dynamics of measurement** as such. When a certain outcome is a desired outcome, people will become opportunistic to reach that goal: Teachers will teach to the test, and students will study to the test. Schools may be tempted to try to play the system in order to attain advantages. For example, if improvement is rewarded, there is an incentive to downplay performance in the first measurement (Hargreaves and Shirley, 2009). In other words, "*when a measure becomes a target, it ceases to be a good measure*" - this general observation is known as Goodhart's Law[44], and it applies in educational settings as much as in business. With big data, the temptation is to try to counter this by measuring more, which does not solve the basic problem, but starts a self-reinforcing dynamic of surveillance (Wright and Kreis, 2014, p.191).

Another challenge derives from the socio-political context of personal data and the consequences that data protection rights have on data sets. As argued above, learners (and, because in schools these are mostly minors, their parents) as a rule have the choice to not have their data included in some collections. That is seen in rules such as the necessity for parents to give consent to their children being photographed (at school fairs etc.). Such consent/non-consent may cover many occasions, or change from one to the next. Parents may opt out of data collection for very different reasons, including political objections to surveillance or religious motives.

Thus, the resulting datasets may be very biased – and it is not even clear which types of pupils are missing, i.e. what the nature of the bias is. However, if data sets are non-representative and, worse, one does not even know in what way they are non-representative, conclusions drawn from them are weak. Worse, such data may lead to conclusions that disadvantage and discriminate against poorly represented types of pupils: for computational analyses of such biases see (Hajian et al., 2016).

By definition, compulsory education (and by extension compulsory education monitoring) violates fundamental rights, and school laws recognise this fact as well as the need for balancing interests.

Throughout history, some parents have objected to their children participating in some school-based activities – swimming lessons, religious instruction, science instruction (e.g. evolution theory), social-science instruction (e.g. homosexuality), visits to Holocaust memorials etc. Depending on the case, such **opt-outs** may, or may not, be granted. With repeated cases, consensus must be reached, a consensus that should be based on **shared social values and/or laws**. For example, European schools today are based on a consensus that all children should learn how to swim, should be instructed about certain contents, and should not be supported in thinking that the Holocaust did not happen. Religious instruction, on the other hand, has become optional in some school systems, or supplemented by the alternative choice of ethics instruction. As a rule, an activity has to be considered an **integral part of schooling** in order to be made compulsory (i.e. not allowing for opt-out). School laws and court cases regulate this delicate balance, in which fundamental rights may be restricted.

This presents data-based educational monitoring with a bootstrapping dilemma: to establish that means M is proportional for reaching goal G that violates fundamental right R, one needs to demonstrate that M is suitable. However, through people exercising their right R, the evidence base for M is degraded. Conversely, if participation is made compulsory, R gets violated from the beginning. Making participation compulsory in the name of other values does not solve, but may aggravate this dilemma: For example, it could be argued that more data collection from all is a measure for inclusion (see the section on Student tracking) – but this inclusion of all into surveillance may well be a false trade-off between privacy and inclusion.

[44] https://en.wikipedia.org/wiki/Goodhart's_law

Anonymous data

A possible solution to this dilemma is to use **anonymous data**, which are not considered personal data and therefore fall out of the scope of data protection.[45] In education monitoring, EU member states (or other political units responsible for education, such as federal states) have taken different approaches at different times. For example, in the PISA study (which requires participation percentages above certain thresholds to be statistically reliable), some have made participation voluntary, while others have made participation in the achievement test mandatory, but participation in the personal-data questionnaire voluntary (Beckmann, 2003). This resulted from an application of the applicable data protection law. The resulting data on academic achievement are anonymous and therefore not subject to the protections of data protection law.

It is much harder to collect and/or create **anonymous** or anonymised data from big data sources, since the uniqueness of a person's data trail increases with the degree of detail of these data. Nikiforakis and Acar assess the uniqueness of the set of browser variables that can be collected from a single click on a web page[46] and other modern methods of device fingerprinting, (Nikiforakis and Acar, 2014).

However, even if this succeeds: Anonymous or anonymised data may still have significant effects on learners' lives. For example, it is conceivable that the results of a classroom study fully based on anonymous data could show a large learning effect of an intelligent tutoring program, and are interpreted as indicating that fewer teachers are needed.

Another risk is the risk of **discrimination** through Big Data Analytics (BDA). Discriminatory effects may result from the analysis of personal or of anonymous data, and while the GDPR explicitly mentions discrimination as a possible consequence of processing that data controllers have to take measures against, an in-depth understanding of both the algorithmic and social discriminatory factors and effects is only just developing (see the following sub-section). Last but not least, the performance algorithms designed to personalise learning may suffer (for example, in the accuracy of the recommendations they give to learners) when anonymous data is used for training the algorithms. This **privacy-utility trade-off** is widely accepted in the data mining community: see (Li and Li, 2009) for a critical review.

Longitudinal data

Longitudinal data collection and storage about individuals can arise in educational monitoring for various purposes and reasons. E-Portfolios are "*student-owned dynamic digital workspaces wherein students can capture their learning and their ideas, access their collections of work, reflect on their learning, share it, set goals, seek feedback and showcase their learning and achievements*".[47] There is the potential for an increasingly detailed audit trail to follow a student almost by the day (Kamentez, 2014). International surveys that build big data include the CEM Centre[48] iPIPS for early childhood education,[49] which covers the following domains: cognitive development; personal, social and emotional development; physical development; behaviour; and, contextual information, although it will not provide internationally comparative data, but instead create "*country level results using scales which parallel those used by PISA, so that contextualisation is immediately possible*".

[45] Pseudonymous data is data that can relatively easily be de-anonymised. The remarks in this section also apply.
[46] https://panopticlick.eff.org/
[47] Cited from http://eufolio.eu/, see also projects such as http://europortfolio.org/ See also the section on Assessment.
[48] CEM provide assessment monitoring at all levels, along with analysis and secure data facilities. http://www.cem.org/assessment-monitoring-systems
[49] http://www.ipips.org/

Such portfolios are by nature more information-rich than traditional certificates and also pupil files. The motivation for this includes convenience (having all one's certificates ready at hand) and opportunity (in electronic environments, it is easy to add for instance an "ideas" space to the spaces for finished works). Opportunity is also related to the reasons for more detailed data in BDA contexts: Big data is often collected opportunistically following availability, speed and cost. If their processing that does not lead to useful predictions and conclusions, there is always the promise that with more data, collected over a longer time, analyses and outcomes will become better.

This can easily lead to **long-term surveillance**. *Regardless* of whether an improvement of descriptions, predictions, and ultimately policies will result from more data, this is *certain* to result in learners carrying an ever-more detailed life-long data dossier with them. At the same time, the same data, aggregated differently, also form detailed life-long data dossiers of teachers and other people involved in learning processes. The longitudinal nature of these data collections poses challenges over and above their level of detail.

Humans go through phases in their development, and this is particularly true for children and adolescents. At various times in their lives, in kindergarten and in school, children are classified, often as: (about which schools have documentation duties) slow learners, ill, afflicted by mental problems, taking drugs, etc. The reasons are manifold, they can derive from external circumstances (such as school personnel planning profiting from more handicapped children), chance, and causality is and maybe cannot be determined. In addition, barely a day goes by without some new report that a personality trait or state (including such categories) can be predicted from yet other data, e.g. social-media data (Kosinski et al., 2013). And these datasets, in turn, will be fed into further prediction systems, such as those predicting study success or employee productivity (Knight, 2015).

In today's school system, most of these categorisations are dropped at well-defined transitions such as those between kindergarten, school, and tertiary education, with teachers and schools exercising discretion over what to put on record. Relatively data-poor dossiers (the grades of a handful of subjects, potentially augmented by verbal transcripts) remain. Even these stifle many an individual's plans in life, e.g. by streaming students into distinct educational tracks (see the section on Student tracking). However, today's school system has a built-in 'right to be forgotten' of many details (such as learning difficulties in kindergarten or primary school). After these transitions individuals get a (relatively) fresh start. (Compare this with offences in penal offences that get erased after a certain period.) This gives people agency. Equipping individuals with detailed life-long learning dossiers could deprive them of the liberties that arise from this forgetting. The wealth of data reduces people to inanimate objects with measured characteristics that predict their futures.

It could be said that life-long learning dossiers resemble a CV, a scientist's publication list, or an artist's portfolio. This is true, but there are two major differences. First, in many circumstances it is allowed or even expected that CVs, publication lists and portfolios get redacted, re-written, and adapted to life circumstances and the context of their use; as long as these modifications do not introduce lies. Second, these dossiers are created by adults who, our society assumes, know what they are doing, both in the life choices they make and in how they describe these. In other words, societies today afford adults some agency in creating their own dossiers, and in the phase of their lives that is dominated by employment or other economic activity, require them to exercise this agency.

Schools, on the other hand, are spaces for minors whom societal consensus regards as not (fully) able to overlook the consequences of their own actions and therefore in need of protection. Going back to the original meaning of 'school', schools are also 'free

spaces', spaces for keeping clear (of e.g. economic pressures).[50] Data collection and processing towards detailed life-long learning dossiers would introduce these pressures in an underhand way and subvert the very idea of a (relatively) free, protected space. (Another potential of big data is a two-edged sword in this respect: the continuous assessment of student performance, as opposed to testing at intervals through milestones (see the section on Assessment). Continuous assessment obviates the need for such milestones and the time and nerves they cost, but it also puts learners under continuous stress, emulates a business environment, and risks eradicating the free space in which errors may be made without sanctions.)

Those with savvy parents may tread paths that avoid the creation of nefarious data, ranging from over-affirmation to corruption, but the naive students and parents may become victims of stigmas assembled over a life-time. Thus, another danger of such dossiers is a deepening of **social divide**.

What is "evidence"?

One of the tasks of this document is to explore the role of big data in developing more effective evidence-based policies. Some research questions the expectations that more data, produced at greater resolution, covering more variables, and analysed using sophisticated analytics, will lead to better policy making. However, in 2014 Gartner, presenting the Education Hype Cycle, warned that "*Big data is at the Peak of Inflated Expectations*" (GARTNER, 2014).

For example, reducing early school leaving can be achieved through a particular set of actions, not just within-school data surveillance activity. However, most educational data mining and learning analytics focus on predictive models, rather than on didactic interventions that really utilise insights, and too often, big data analytics and attempts at evidence-based policies neglect context, risking the creation of a "*data-centric and data-intensive capitalism*" where citizens have no control over data (Morozov, 2015).

Claims that big data has huge potential for learning and monitoring translate to a number of assumptions, which are summarised as follows. Firstly, there is lots of data. Second, there is lots of data that is easier and cheaper to collect than data used in traditional education monitoring. Both of these are generally true, but they may reduce expediency. Third, the data tell us something. Fourth, that it is easier, cheaper, and/or faster to analyse the data than it is to analyse traditional data: that may bypass reading student texts and instead have machines read and grade them. Thus, taken to its extreme, the assumption is that we may bypass the costly testing of learning outcomes by PISA-like questions and replace it by logs and sensor readings, from which machines can reliably predict learning outcomes (or at least do the assessment, see the section on Assessment). It is necessary to realise that the last two are assumptions that require empirical evidence (see for example the challenges of Learning at scale described in the section on Assessment). The mere claim of "potential" does not constitute evidence.

An example of this concern from another context is the debate around the benefits of biofuels as a way to reduce dependency on fossil fuels. Saltelli and colleagues examined the ways in which large volumes of data were used to build evidence-led policy, but concluded that there was too much "*use of statistical indicators and mathematical modelling used outside their semantic context as an element of obfuscation and distraction from uncomfortable knowledges*" (Saltelli and Giampietro, 2016).

Effective use of analytics, and the ability to critique them within a social context, will be more important as the big data approach starts to use artificial intelligence and adaptive learning approaches to deliver highly personalised learning environments (Gros, 2016,

[50] http://www.etymonline.com/index.php?term=school

Williamson, 2015b). This will raise further issues regarding who is liable in cases of malfunctioning or misuse (Lynch, 2017, Harris, 2016).

Institutional policies and competencies

There is a rapidly changing landscape of digital threat. Eurostat has reported that across the EU in 2015, 25% of internet users experienced security problems, ranging from viruses, hacking, financial problems, or "*children accessing inappropriate websites*" (ESTAT, 2016a). Against that, institutional IT procurement policies present a security paradox, because they can be slow when centralised, or fast but heterogeneous when procurement is delegated. There are risks that the speed of investment decisions lags significantly behind security threats. Innovations such as the latest router protection for integrated (internet of things) homes (Ward, 2017) may take time before they are marketed and used, but they are often designed to deal with existing problems.

Combined with software and hardware security risks, there are risks of uneven knowledge and competences within school staff – everyone who uses IT facilities from managers, teachers, to assistants and office staff. Cybercriminals play on the uneven knowledge in the same way that fraudsters do via email scams. In the UK, late in 2016 fraudsters were phoning head teachers and administrators, claiming to be from the education ministry, and informing them that important and sensitive information would be sent to them. A document would contain ransomware "*that once downloaded will encrypt files and demand money (up to £8,000) to recover the files*" (NFCRC, 2017).

When big data analytics are used, further risks arise from the lack of understanding of these methods and their issues by school staff, who are usually not data-mining experts.

A further danger arises from staff, even if privacy-aware and well-intentioned, perceiving data protection as a burden. For example, to increase equity, the German *Land* Berlin offers poor families a specific subsidy to enable their children to take part in cultural activities. For data-protection/privacy reasons, teachers may not ask directly whether a child is part of this programme, but must consult (paper) files in a cumbersome manner. Only in this way can they find out how much money they have to pay themselves in an upcoming activity (such as a visit to the cinema in class), to be reimbursed later. Teachers thus also carry financial risks privately. In many cases the consequence is that either rules are broken (the information is requested and shared otherwise) and/or that fewer activities are carried out.[51] Thus, it is not necessarily a lack in understanding, training, or expertise that endangers data protection, privacy, and other rights, but poorly designed system interfaces and the lack of recognition that privacy and data protection, done properly, are work activities that demand time.

Socio-political effects in interaction with data protection and privacy

Shifts of expertise and power

If programs diagnose learning outcomes, if programs predict future educational achievement, if programs thereby determine who gets recommended for what, the locus of expertise and power shifts from teachers and school administrators to – depending on one's perspective on machine intelligence – programs and/or their designers and/or the institutions that perform these BDA. These shifts in decision-making pose a range of risks including **de-skilling** (known from other domains; see (Condliffe, 2016, ECONOMIST, 2017) and **transfers of public responsibilities and powers to private actors that are not subject to democratic control** (Taekke, 2011).

As a consequence, the fairness and transparency of processing (a key principle of the GDPR, see Article 5) may be endangered. In the domain of criminal justice, where similar problems are encountered when, e.g. future criminal behaviour is predicted and parole

[51] Source: personal communication with a secondary-school teacher and with several teachers in tertiary education.

decisions are recommended, two problems are clearly emerging from the various ways in which BDA are opaque. There may be a lack of transparency and accountability because first, algorithms and software are proprietary, and second, even if open to inspection, algorithms may be "*non-interpretable*" (Lipton, 2016). Such non-transparency can have many adverse effects. One that is currently being discussed is unlawful discrimination, see (Barocas and Selbst, 2016, Berendt and Preibusch, 2014), and the sections on educational efficiency and equity and on student tracking.

A focus on algorithms hides another challenge of such outsourcing: the availability of (other) data that the data processor can link to the newly acquired data. In principle, BDA algorithms can be developed in-house (a strategy currently employed by some authorities for predictive policing[52]), which alleviates the problem of proprietary algorithms. Software can in principle also be deployed in-house, which alleviates the problem of data leakage. However, with the current trends to service-oriented architectures (McLellan, 2016) and towards industry concentration (Lynn, 2017), it is more likely that, big vendors will control both software and data, and therefore opportunities for further linkage and profiling.[53] With concentration, in turn, the risks of large-scale security breaches increase.[54]

Data "ownership" and control

CMS and SMS systems increasingly store big data beyond the physical borders of a school. While the systems may be intuitively easy to use (for example, clear menus and effective user training), they are extremely complex in their software design and their hardware configurations. Teaching staff may focus more on the teaching and learning functionalities, rather than understanding the underlying data storage and security functionalities. Furthermore, there can be issues where schools use Dropbox for student work, since this may breach privacy rules because it stores on the cloud (Kelion, 2015).

There are security challenges where schools allow or instruct students to 'bring your own device' (BYOD) (NMC, 2016), whose opportunities include seamless work. While BYOD may save money in terms of hardware procurement, it may increase security risk, with software and device proliferation on multiple devices, and leaky individual firewalls where the devices have different (or no) security software: "*it increases overheads such as internet infrastructure, software licensing and technical support*" (Bird, 2016). In addition, BYOD may present inclusion challenges, because not all students have home environments that are internet-linked with devices for the students.

Not only does data previously owned and controlled by schools move outside its previous confines. The same may happen to "student-owned" data, even if ePortfolios market their products in this way. Challenges to ownership and control arise because data relating to the students, which can be highly detailed and distributed across both administrative sources, and commercial IT platforms, where the data may be stored on the cloud and beyond the jurisdiction of an education system (COMMONS, 2014).

While the word 'ownership' has been used with caution here, the challenges are actually more about effective control and usage rights. First, the spirit of European data protection is precisely *not* focussed on who owns personal data, but about these being protected no matter who owns them, for example by enabling individuals to always control data about themselves. Treating personal data as a property could entail the ability to sell not only one's data but also the rights on them (which would weaken

[52] Cf. an overview for Germany: https://blog.pilpul.me/wo-predictive-policing-eingesetzt-wird/

[53] For example, two of the previously mentioned platforms integrate Microsoft Office 365 (EUFolio) resp. Microsoft Office Online (Smartschool). See the outlook of the section on Assessment for further issues of monopolistic structures.

[54] Cf. the recent purchase of Lynda.com by LinkedIn (Owsinski, 2015, Kapko, 2016) and the subsequent data breach (Hacket, 2016).

individuals and increase social divides), whereas a fundamental right cannot be waived: for a discussion and references see (De Wolf et al., 2016).

However, even if many provisions of the GDPR aim to ensure such control and at least reduce the commodification of personal data, control can also be reduced by convenience (e.g. having stored data in a format proprietary to the platform and/or the software vendors they work with), poor interfaces for interoperable formats and little expertise in using them, as well as usage rights that individuals or institutions grant to service providers, often unwittingly (see for example the YouTube terms of service and the widespread use of YouTube in educational institutions).

Agency

Including and beyond data ownership and control, what is the agency of the participants in the learning process? This discussion will focus on students, but it should be taken into account that similar arguments apply to other stakeholders such as teachers, about whom big personal data are collected and whose range of actions may be constrained by educational software and educational monitoring.

Clearly, students should not just be data subjects. However, which competencies are expected of them, and what do we expect them to take charge of? What degrees of freedom do they have? For example, how can students determine the 'best' pieces in their portfolio in an eKool-like system, and how can they adapt such decisions later? Are these selections logged, who can see the logs, and what does this mean for the freedom of choice?

How is it possible to tread the thin line between respecting stakeholders' autonomy and burdening them with additional tasks just to manage their identities in the face of a growing data footprint (responsibilization, see (Shamir, 2008))? Is 'giving people choices' really respecting their agency, or just a neoliberal illusion of consumer choice (Jones et al., 2013, p.153)? Is a pervasively digital public environment engendering democratic behaviours, or is it turning people into de-politicised, passive consumer citizens, as has been argued for the Estonian system (Björklund, 2016)? Such tendencies may be furthered by one of the features that big data proponents regularly tout as one of the key advantages: individualised learning. Individualised learning may crowd out a sense for the importance and power of collective action, in learning as in other civic activities.

How is it possible to tread the thin line between respecting students' autonomy and the need to protect them as vulnerable? For example, is it "*in the best interest of the chil*d" (UN Convention on the Rights of the Child, Article 3) to insist on parental consent (which the GDPR now does up until a default age of 16) or to let a 13-year-old decide for themselves (as under US law, which influenced earlier versions of the GDPR and is now an option for Member states)?[55]

Context and boundaries

Arguably, the object of the EU fundamental right to privacy is a private space that can only be protected if it is bounded. Violating these boundaries can happen through data-related activities, but data need not be involved. (For example, receiving an SMS from a teacher on the weekend may be a privacy violation in this sense, even if no personal data are involved.) The discussion of this last challenge, therefore moves beyond data protection.

Traditional social and institutional life was characterised by many boundaries and social contexts that were kept clearly separate. These boundaries are disappearing, and

[55] For examples of this debate, see https://iapp.org/news/a/will-gdpr-move-age-of-consent-to-16/, http://blogs.lse.ac.uk/mediapolicyproject/2016/03/01/eu-general-data-protection-regulation-article-8-has-anyone-consulted-the-kids/

contexts "collapse" (boyd, 2008). This problem has been discussed extensively with respect to social networking platforms, in which the broadcasting of information to 'all friends' is causing regret and many other problems: for example see (Gao and Berendt, 2014). The phenomenon can be observed in educational contexts too. Examples include students' use of their integrated devices (and the software on it) as BYOD at school, for their homework, and for their social life. Boundary dissolution is also touted as a key productivity feature of CMS/SMS, in which administrative and pedagogical data is linked. Sensors and embedded systems make data collection much less conspicuous than the use of desktops, laptops and even earlier mobile phones. Thus, students, teachers, and other persons involved in education use a wide variety of software and resources from "outside the school walls" and also leave digital trails there.

All this leads the potential for information to be "*captured by a surveillant assemblage devoted to the disappearance of disappearance*" (Haggerty and Ericson, 2000). This situation, where online information about individuals could lead to potential harm, contributed to the creation of the "right to be forgotten" (more accurately, the right to the erasure of some data) (Commission, 2012a) and its inclusion in the GDPR. However, with the proliferation of resources such as the Internet Archive[56], and the ability to copy and share information, it is increasingly difficult to 'disappear' online.

In addition to having more opportunities (and therefore more responsibilities) over their own data and lives, all those involved in the educational process also attain more opportunities and responsibilities over others. Privacy is mutually constructed, and individuals cannot have their privacy without respecting that of others.

This creates new dependencies and vulnerabilities for individuals and organisations. While convenience suggests that educational institutions and their members continue eliminating context boundaries, social and political analyses emphasize their importance (Turkle, 2011). The data protection principle of purpose specification and limitation is an important boundary keeper between what data has been collected for.

Implications for institutional, national, and EU policy

The study was asked to consider developments may be expected in the coming 10-30 years, where the focus does not lie on technological developments (this is the topic of other sections in this report), but on developments regarding privacy and data protection. The societal discussion on these rights will continue to evolve, although the main tropes have existed for a long time. The GDPR was the result of decades-long development of data processing and privacy and data protection debates. It is a complex and far-reaching law, and it requires ambitious new processes from data controllers and processors. There is much hope that it will present a viable legal framework for as long as its predecessor did, i.e. at least 20 years.[57]

Institutions from local level schools up to the supranational level of the EU will face substantial challenges just trying to put the requirements of the GDPR into practice (Tsormpatzoudi et al., 2016), and a serious commitment to doing so can improve the protection of citizens' rights substantially. This subsection therefore concentrates on exploring what this would entail. It starts at the level of institutions and national decisions, investigate education monitoring choices, and then sketches further policy avenues at the EU level.

[56] https://archive.org/index.php - in January 2017 storing 279 billion web pages. It takes 'snapshots' of web sites, so that even if information is deleted 'today' it still exists in previous versions of a site, and they are publicly available at no cost.

[57] Source: Discussions at CPDP 2017, one of the major policy conferences in the area in Europe, see https://www.youtube.com/user/CPDPConferences.

Institutional and national policy avenues

GDPR building blocks: data minimisation, data protection impact assessment, data protection by design

For every planned data collection, software or hardware deployment, etc., it is imperative to take recourse to a fundamental principle of European data protection law: data minimisation. Can the same (e.g. learning or monitoring) effect be achieved with the help of less data? This amounts to a necessity test for proportionality testing.

Suitability tests for proportionality testing must rely on scientific evidence, but care should be taken to not be blinded by an overly de-contextualised "scientific" method. Saltelli et al. (Saltelli and Giampietro, 2016) advised that the statistical analysis should be accompanied with a more "socially robust alternative" of quantitative story-telling, which "*involves a participative and deliberative analysis of the quality of proposed or available policies and narratives on governance*".

Once a teaching and monitoring intervention has been (re-)designed to work with less data, the proportionality test still has to be applied: Is the big-data method suitable, necessary, and reasonable in relation to how much it interferes with the right to data protection, and with the rights and freedoms of individuals that may result from data processing: privacy, freedom of speech, and the enjoyment of the liberties of a democratic society?

How can a proportionality test be initiated in a principled way? The GDPR includes the mandate for a very helpful method that answers some of these questions: a **data protection (or: privacy) impact assessment** (PIA). This involves the identification of key stakeholders and their interests in a proposed new technology or method, as well as how their rights could be affected. Which information flows, from where to where and how? What are the roles of the stakeholders in providing, disclosing, collecting, using and sharing the information and the purposes and outcomes of analytics? Based on the state of the art in **privacy-enhancing technologies** and processes, how can the negative effects be mitigated?

This last step amounts to applying **data protection by design**, which is however not limited to deploying certain technologies, but also involves organisational measures (Danezis et al., 2015). In addition to more traditional PETs, this should also utilise current algorithmic and procedural developments for making processing discrimination-aware, transparent, and accountable.[58]

Institutional rules, guidelines, training, and support

The recommendations of the previous subsection assume the existence of a high level of competencies and resources in educational institutions. Approaching such an ideal state, however, requires an enabling institutional frame.

To minimise the risks of data protection and privacy breaches, a strong system-wide policy is essential, providing both rules (security, data protection, safeguarding of children etc.), guidelines, training, and support. One example is the approach of the German Data Protection Authority of the State of Schleswig-Holstein (ULD). This ranges from clear policy guidance, such as teachers not communicating with students via private emails, or through social networks and Apps such as WhatsApp – and laying out clearly the consequences for breaching the policy. Such an approach both protects students from risks, and also maintains a clear and secure audit trail of communication within the authorised school system.

The ULD approach includes twice-yearly training for head teachers, training for administrative staff in schools, mandating that encrypted USB sticks are used, along with

[58] See e.g. http://www.fatml.org

strong security procedures for home use of data by teachers (previously a paper process, where teachers historically took written assignments home for assessment), strong data protection and privacy principles on school web sites, and where students are able to access learning platforms at home "*the use of the WLAN is only possible with an access code*" (ULD, 2012).

Such approaches need to be reviewed regularly to keep up to date with technological and market developments. An example are the instructions and enforcement of data protection where data is to be stored on the cloud. To the extent possible, clear and authoritative check-lists should be given, e.g. (DfE, 2017a).

At EU level, ENISA commissioned a study on training and support needs for Network and Information Security (NIS). Proposals included the development of a Europass document for NIS skills, better education and continuing education of teachers (being the key multipliers to students and other staff), develop scenarios for data protection, and develop an NIS MOOC (Berendt et al., 2014).

Parents are also part of the education process, and need to provide guidance to their children, but a UK survey has noted that, even in a highly-connected society, there were big disparities between the IT skills of children and parents. Parents were much less competent, and therefore much less able to engage with their children to develop the competences, awareness and skill to be 'digital citizens', and often have fears that "social media hinders or undermines moral development" (Burns, 2016c).

Since data protection impact assessments and data protection by design are only becoming mandatory with the GDPR, currently there is limited experience on how to do this and even more so how to teach it to decision-makers. No doubt, the GDPR will lead to the growth of a new consulting ecosystem. Public-private partnerships could help develop professional software while maintaining data protection standards.[59] However, it is important – especially in schools as educational institutions – that at least a basic understanding is also created and maintained locally. There is encouraging feedback and results from the development and deployment of a simple form of PIA teaching in educational contexts (Tsormpatzoudi et al., 2016). A recommendation is to develop such efforts into regular offers for schools and other educational institutions.

To counter the trend towards shifts in decision making and power, it appears particularly relevant to turn PIAs into a democratic exercise. Thus, stakeholders should not only be modelled, but integrated into the design and development of systems possible (co-design). This will present additional challenges and opportunities in school settings, in which consent and votes of affected individuals (students) may need to be supplemented or supplanted by those of their legal guardians (parents). At the same time, it may vastly enhance the civic education of students and help them learn about their rights in a democratic society first-hand. See (Berendt et al., 2014) for examples of teaching these values at secondary schools see. In addition, more participation in system design and development is needed to avoid that privacy and data protection are perceived as a burden and therefore neglected or, maybe worse, regarded with cynicism.

Integration with curricular contents

As data controllers, schools and school authorities are responsible for *safeguarding* their members' privacy and data protection rights. As educational institutions, schools are responsible for ***teaching*** about privacy and data protection as rights and as societal phenomena, just like they are responsible for teaching about freedom of speech and other fundamental rights. More and more institutions are recognising the key importance of such teaching, and more and more materials and courses are being offered that also

[59] An example is the UK Gradintel (http://gradintel.com) data platform.

profit from the interdisciplinary potential of the topic and its interestingness for learners (Berendt et al., 2014).

Privacy and data protection should become a core element of secondary (and maybe even primary) education, education monitoring should be discussed at least outlined, and all opportunities for promoting these issued should be taken in the teaching and learning environment. In particular, this means that schools and school authorities set an example of good data-handling practices, do not contradict the contents of what they teach by their administrative practices (e.g. using communication services with dubious data-handling practices, just because it is convenient), and offer transparency and democratic participation to the extent possible.

All school members need to be provided with the knowledge and skills to understand the balance between rights and obligations, as well as being able to provide informed consent relating to their own information. Schools and authorities should contribute to safeguarding boundaries by technical and organisational means, and by appropriate training and education for their staff.

What data for education monitoring?
An inspection of typical indicators used in education monitoring today (Eurydice, 2017a) (Eurydice, 2017a), and a consideration of the data that underlie them, can be summarised as follows:

- The finest granularity is at the level of demographics of a pupil (e.g. socio-economic status), sometimes with longitudinal data (school progression). In other words, this is the level of today's student registers;
- Finer-grained data such as PISA test results are (at least in data-protection-conscious countries) anonymised;
- For teachers, only "input" variables (such as whether they received a certain training or not, whether they get reduction in hours) are used, their "output" is at most measured by school performance – where it is, for big data purposes, aggregated at school level;
- Student-centric input data are aggregated at the level of class or school (e.g. do schools get language support for students with a mother tongue?);
- Behavioural data for neither pupils or teachers are used;
- Indicators are collected on the basis of research findings. Specifically, *factors* (such as socio-economic status or the language spoken in a student's home) are related to *interventions* (such as additional staff that schools receive or language support they offer) because there is *evidence* that the intervention improves the target *outcome* (e.g. achievement in basic skills). An intervention may also target a factor (e.g. language competencies).

Should more measures be included in indicator lists, should more data be collected, or more data be re-purposed; For example, data relating to curricula, or student-level analytical data such as that available in eKool, or even more fine-grained data, including of the operational kind, such as sensor traces from educational software.

Is it argued here that such extensions must be subjected to in-depth **proportionality tests**. In the following, the term 'use of data' denotes any collection, re-purposing, and processing activities. The following questions need to be asked (before laws are considered that prescribe such uses of data):

1. Does the proposed use of data serve a *legitimate aim*? This should be a target outcome germane to education monitoring (such as skill levels of a population). Other big-data-amenable aims have questionable legitimacy in this context. For example, even if skills forecasting at a society-wide, aggregate level is undoubtedly a social good (see

also the section on Skills forecasting), should every individual have a skill profile that can be matched to a current labour market demand? It is important that European views on education should continue to view it as a public good (Daviet, 2016) and not only a job-market instrument (in addition, such attempts at (over-)fitting to current economic needs are bound to fail).

2. Is the use of data *suitable* to achieve this aim? If the data is factor-related, is there *evidence* that this factor contributes to statements about interventions and outcomes? If the data is intervention-related, is there *evidence* that this intervention contributes to outcomes? In other words, before large-scale and mandatory data collection is considered, scientific studies should have been carried out, with the necessary safeguards for consent and data protection, that provide evidence of the usefulness of the data in the given context.

3. Is the use of data *necessary* to achieve the aim? If there is evidence that the same aim can be reached with less data, it is not necessary. If there is no such evidence, it should at least be sought for.

4. Is the use of data *reasonable*, i.e. have the consequences on fundamental rights been assessed against the objectives pursued (balance of interests)? Investigating this question elaborates on the motivation of indicators provided by Eurydice. The following passage refers to factor-related data, but the same reasoning should be applied to intervention-related data.

The Commission (Eurydice, 2016e, p.6) explains why they limit the factors on which they report:

> "*The importance of out-of-school factors, including students' socio-economic background and the educational level of parents or the language spoken at home cannot be overstated. Significantly reducing the proportion of low achievers, therefore, would require a combined approach that simultaneously targets a range of factors both in and out of school. The following 2016 structural indicators, however, concentrate primarily on factors that can be directly influenced by education policies*."

This reasoning can be extended to a big data context. There are undoubtedly many further in-school and out-of-school factors that are important for educational attainment. For example, these could be behavioural aspects such as the number of hours a student spends online in their free time, or spends in different types of virtual environments online in their free time. There is some empirical evidence on some such factors (see the quotation from (OECD, 2016a) in the Introduction of this report). Regarding other factors, there may not (yet) be any evidence and instead only hunches based on the easy availability of certain data with "big data" hard- and software (an example are the manifold scores generated by fitness-tracking devices). Many other trait or state variables may have an influence (e.g. IQ, religion, political or sexual orientation) and may be inferred easily from available big data. For work about the prediction of such variables from Facebook Likes see (Kosinski et al., 2013).

Some of these factors may be targetable in principle (i.e. influenced, even if indirectly, by education policies – number of hours online or number of steps walked in a day are examples). Others, as in traditional indicator settings, may hardly be targetable by education policies (for example the education levels or socio-economic status of parents). Yet others may be targetable, but education policies that try to influence them would clearly be unethical (e.g. religion, political conviction, or sexual orientation).

Not only the attempt to influence some factors, but also their collection and processing may be unethical or even illegal. Both targetable and non-targetable factors may correspond to sensitive data per se, and/or their collection may interfere in the private lives of students and teachers. All the above examples were chosen to illustrate this.

Thus, both the degree of targetability and the possible infringement of fundamental rights must be taken into account in deciding whether the use of data is reasonable. It is important that educational authorities and the EU as a whole honour their obligations to protect every citizen's, and in particular children's, privacy and data protection, and do not jeopardise this goal by an inappropriate normalisation of electronic surveillance.

EU policy avenues

The EU has a key role in applying fundamental protection across Member States and building knowledge and practice. The lessons of the past tell us clearly that the future will see privacy and data protection rights being challenged ever faster (speed), in greater detail (more big data and more analytics), and more pervasively than before. The GDPR is a comprehensive legislation that contains many elements with the potential to accommodate technological and social changes, and these elements are far from trivial. Many concepts and rules remain to be interpreted, and national derogation rights will be used in diverse and yet-unknown ways. A new ecosystem of consultancy will develop to help build the technical and organisational means for safeguarding data protection, which are referred to in several places in the law. Such consultancy should not only be accessible to the well-off or well-educated; care will have to be taken to avoid new inequities.

One motivation for changes in the GDPR is that the old legal framework, the Data Protection Directive, already contained key principles and rights, but lacked enforcement (Article29, 2010). There is ample opportunity for showing the teeth of the GDPR even now, with projects that are well-intentioned but that raise too many questions. One example is the UK National Pupil Database, which makes data available to third parties with relatively low thresholds.[60] Critics point out issues such as the lack of knowledge, let alone consent, by the 20 million people whose personal data, including sensitive data reside in "*one of the richest education datasets in the world*", and the lack of meaningful barriers to abuse[61] as well as the questionable value of the data.[62] Cataloguing such examples, developing ways of making them compliant with data protection principles, and enforcing the requisite changes, will be a long, arduous, but indispensable part of a strategy for effectively protecting individuals in the face of educational technologies and education monitoring.

This will need to include collaboration across regulatory areas. For example, the risks associated with data held by large commercial players can neither be countered by informatics or educational experts alone, and also not by data protection experts or authorities – they need a joint effort with (at least) competition law. The EU Data Protection Supervisor has recently formulated this as a core policy objective (EDPS, 2017).

Laws and regulations alone are not sufficient. The more that frameworks of **confidence and trust** can be built into the education big data environment, the more the data systems can be integrated and the data used for real-time and robust monitoring across and between systems (Lane et al., 2014). Of course, that also requires attention to important statistical aspects of data structures, standards, metadata, anonymization, and access control. But, research recommendations show a need for frameworks that generate good practice and trust, and it needs people, businesses, organisations, and administrations to have the skills and knowledge to use data effectively, and to be clearly accountable for failures (Prinsloo and Slade, 2014). If those are in place there is the potential for a virtuous circle to operate where greater security and trust generates greater understanding and consent from the data subjects. Conversely, failures in

[60] https://www.gov.uk/government/collections/national-pupil-database
[61] http://defenddigitalme.com/
[62] http://schoolsweek.co.uk/?s=%22national+pupil+database%22

security, abuse or misuse of data, and well-founded concerns as well as less-well-founded fears all contribute to reducing trust and diminishing consent.

The EU can further support the ongoing implementation and operationalisation of the privacy and data protection framework through its use of 'soft power' mechanisms using the Open Method of Coordination[63] (OMC) DG EAC supports Member States in developing their education systems, in particular for schools[64] through analysis, information sharing, building capacity and good practice. The collaborative initiative European Schoolnet, involving 30 ministries of education in Europe, has an Academy[65] which makes available courses such as ICT Infrastructures in Schools, and approaches to collaborative teaching and learning. It is linked to the Future Classroom Lab[66], which links policy-makers, industry actors, teachers, and a wide range of education stakeholders. Such an action can help to sensitise education systems to forthcoming developments, at the same time sensitising the technological innovators about the regulatory and ethical challenges.

In a recent review of US big data education developments in the context of the EU, Yoni Har Carmel made some clear policy recommendations which emphasise the foundations provided by effective regulation. Potential policy risks need identifying, clear boundaries need setting between public and private data 'spaces' in digital learning environments, clear limits need setting on who can access (access control) and use information, and dialogue, awareness, and knowledge sharing can help to build trust (Har Carmel, 2016).

Such actions are confirmed in a broad-ranging study of the big data landscape, ENISA (the European Union Agency for Network and Information Security) emphasised that privacy/data protection must be 'designed in' at all levels, ranging from technology devices, to software, algorithms, and (through education and skills) into people. Their major recommendations included:

- **Collaboration and dialogue** across all actors in the big data landscape "*to define how privacy by design can be practically implemented (and demonstrated) in the area of big data analytics, including relevant support processes and tools*";
- Ensure that **privacy policies** are applied automatically;
- Enhance the **consent process**: "*the very idea of consent needs to be reinforced with new models and automated enforcement mechanisms*";
- Develop better **awareness and promote** effective use of PETs "*privacy enhancing tools for online and mobile protection*";
- Develop a **Commission-wide** "*coherent approach towards privacy and big data*". (D' Acquisto et al., 2015)

The privacy- and data-protection-related actions taken by the Commission have been shown as extensive, but taken as a whole they do not have a clear focus in the context of education systems and big data. The High-Level Group on the Modernisation of Higher Education recommended back in 2014 that there was a need for national authorities to develop digital skill competency frameworks, to integrate them into professional development for teachers, and that training in "*relevant digital technologies and pedagogies*" should be available to all teachers (Commission, 2014b). While the ET 2020 Strategy has a Working Group on Digital Skills and Competences[67], its remit is specifically related to "*the development of digital skills and competences at all levels and stages of learning*", with some activity related to Learning Analytics and data in

[63] http://ec.europa.eu/culture/policy/strategic-framework/european-coop_en
[64] http://ec.europa.eu/education/policy/school_en
[65] http://www.europeanschoolnetacademy.eu/
[66] http://fcl.eun.org/
[67] https://ec.europa.eu/education/policy/strategic-framework/expert-groups/digital-skills-competences_en

education. There clearly is a role for a more coherent focus on big data and education systems at the EU level.

At the same time, any policy must consider not only coherence and similarity when envisaging different scenarios in which big data could develop in education systems. As current indicator systems illustrate very well[68], the EU is diverse in its education systems at many levels, with many decisions taken to accommodate national or regional specifics (such as large minorities, federal structures, etc.). Arguably, especially wide-ranging choices concerning digitalisation of public life depend on the cultural context and also contribute to it.

For example, Björklund has argued that the pervasive eGovernment system employed in Estonia would only be possible there and shapes national identity there, in a way that reflects the country's recent history and values (Björklund, 2016). The education monitoring platform is a part of this system, and it appears likely that this would not be acceptable in other countries. Cultural factors might lead, even if such systems were mandated, to under-use and also deliberate acts of subversion. European countries differ in their views (and laws) on education, and they differ in their views (and laws) on privacy and data protection. This diversity should be valued as a source of cultural richness and continued democratic debate, rather than stifled by overly monolithic data infrastructures and policies.

Looking ahead 10-30 years in the context of privacy can be particularly contentious. Predictions could range from a dystopian surveillance paradigm, where students are pervasively controlled, to a scenario where the big data tools are used democratically and in a participatory way that perfectly balances the needs of all stakeholders. Clearly, neither extreme is likely. Instead, a policy look-ahead can consider how the EU further strengthens the checks and balances that are already in place, and how they can be developed to be better able to cope with the immediacy (time) and granular scale (individual) of data produced in particular by learning platforms.

Data collection and analysis may be governed by what is technically feasible (e.g. comprehensive sensing, real-time data analysis and purely algorithmic decision-making) and economically plausible (e.g. a missing, poor, or one-sided evidence base concerning efficacy, lack of data post-processing), and it may be motivated by short-term economic objectives (e.g. fitting learners to jobs) and means (e.g. handing over educational data collection, analysis and decisions to the private sector). In such scenarios, there is no time for a 'reflective' consideration of legal implications such as breach of privacy. Instead, privacy and data protection need to be designed into the process from the outset.

In conclusion, current developments offer a unique opportunity for safeguarding European values and fundamental rights in the deployment of big data technologies for education in the next 10-30 years. The GDPR was built with a view to the next decades, and it sketches techniques and processes for protecting personal data and individual rights and freedoms affected by data collection and processing. To enable the radical changes that big data makes possible in the area of education and education monitoring, these changes should be made on a solid foundation. Through data protection by design and the other techniques and processes that have been discussed around, and mandated by it, the GDPR can provide such a foundation, which can then provide more trust and credibility for big data analytics, where those whose data are being processed could be more aware that EU-level checks and balances are designed into the systems.

[68] https://webgate.ec.europa.eu/fpfis/mwikis/eurydice/index.php/Publications:Structural_Indicators_for_Monitoring_Education_and_Training_Systems_in_Europe_2016_%E2%80%93_Thematic_Overviews

Educational efficiency and equity

Introduction: efficiency and equity in education

This section uses the two definitions of the Commission in the 2006 Communication "*Efficiency and equity in European education and training systems*":

> "***Efficiency*** *involves the relationship between inputs and outputs in a process. Systems are efficient if the inputs produce the maximum output. Relative efficiency within education systems is usually measured through test and examination results, while their efficiency in relation to wider society and the economy is measured through private and social rates of return*". (Commission, 2006a)

Efficiency focuses on the relationship between inputs and outputs. A system or process is efficient if a certain input results in a maximum output, or if a certain output is obtained from minimum input (Wößmann and Schütz, 2006). There are two main aspects. First, there is the **efficient allocation** of resources and particularly, balance between different kinds of resources. For example, a balance between the number of teachers per students and highly qualified teachers, or between teachers and whiteboards, or between whiteboards and computers. Secondly, there is the **efficient use** of these resources, making the best use of each particular resource (Wößmann and Schütz, 2006).

Equity was defined as:

> "*The extent to which individuals can take advantage of education and training, in terms of opportunities, access, treatment and outcomes. Equitable systems ensure that the outcomes of education and training are independent of socio-economic background and other factors that lead to educational disadvantage and that treatment reflects individuals' specific learning needs".* (Commission, 2006a)

Equity was re-emphasised in the Rethinking Education Communication of 2012, which noted that while there are significant opportunities to use technology in improving "*quality, access and equity in education and training*", it has been up to Member States to decide what to use, how to fund it, and to balance efficiency with "*equity and access*" (Commission, 2012b). The EU therefore considers systems as being equitable if they first ensure that the outcomes of education and training are independent of socio-economic background and other factors that lead to educational disadvantage, and secondly, that treatment reflects individuals' specific learning needs (EP, 2007).

Balancing policy priorities and resources: Making an education system both efficient (particularly balancing the demands and costs with the supply of funding) and equitable (for example, ensuring that all learners are provided with a quality education, irrespective of their needs) is a significant policy challenge. At one end, there are finite financial resources, and achieving equity is not simply a matter of injecting uncontrolled amounts of funding into a system, since the way a system 'performs' (quality of teachers, quality of teaching pedagogy and content, learning technologies etc.), will influence the ways in which resources are applied.

At the other end, there are the needs of individual learners that are often complex to identify (for example where the family and health circumstances of a child change, requiring multiple agency inputs), and then for the needs to be resourced.

Modernising the education systems: Acknowledging the complexity of a system that is able to incorporate equity and efficiency, there has been a range of EU policy actions relating to efficiency and equity. These have included a focus on digital inclusion[69], on

[69] https://ec.europa.eu/digital-single-market/en/digital-inclusion-better-eu-society

"*Opening up Education*" (Commission, 2013c) across lifelong learning pathways, building open education resources[70], within the ET 2020 Strategy working groups and activities in areas such as digital skills and competences[71], and the reform and modernisation of education systems.[72]

Therefore it is important to focus on efficiency and equity as a complex interaction of processes, and the EU has undertaken initiatives such as supporting the development of digital educational organisations (Commission, 2016b, Inamorato Dos Santos et al., 2016), in reviewing teaching practices and policies (Eurydice, 2015b), and in upgrading education infrastructures (EIB, 2016).

Understanding complex interrelationships: There is not a direct causal link between particular interventions to make things more efficient or equitable. For example, the introduction of teaching technologies could result in efficiency savings, but the OECD warns that PISA tests have shown that students who use ICTs heavily in their learning can have worse learning outcomes (Schleicher, 2015a). Consequently, integrated approaches are needed to enable ICTs to help deliver equity and efficiency, particularly where there are well-trained and innovative teachers, supported by strong school leadership (Schleicher, 2015b).

Starting early with education: The stage at which schooling starts can affect how equity is achieved, and research is clear that early childhood education and care (ECEC) can significantly overcome early equity, by overcoming early socio-economic disadvantages (Council, 2011, ECONOMIST, 2016f). Unless the preparatory work is undertaken at the early stages of education (ECEC) innovation and creativity at later stages will not be as effective as they could be (Council, 2015a). Good education improves social mobility and can produce strong positive socio-economic outcomes (OECD, 2010, Commission, 2013b).

Understanding individual needs of learners: Improving equity within education systems, for example through improving access and outcomes for all students, and particularly those from low-socioeconomic backgrounds, can have a significant economic impact (Commission, 2016a, OECD, 2012). More recently there has been a focus on achieving equity in education systems to reduce the risks of radicalisation of young people (Council, 2016a). Achieving equity is a strong focus for UNICEF in poorer/developing countries, particularly overcoming social, sectoral and infrastructure barriers that inhibit equity (UNICEF, 2010, UNICEF, 2013)

Learning from comparative assessment: Central in the comparative monitoring of equity and efficiency have been the testing programmes initiated by the OECD and by the International Association for the Evaluation of Educational Achievement (IEA). To overcome the complexity of building monitoring data internationally from diverse education systems, they use a process of standardised testing:

- The OECD Programme for International Student Assessment (PISA[73]) has been undertaken in 2000, 2003, 2006, 2009, 2012, and 2015. The 2015 wave covered 72 participating countries and over 500,000 students took the tests which over two hours covered science, mathematics, reading, collaborative problem solving and financial literacy.
- The IEA[74] undertakes international comparative assessments in over 60 countries, assessing student's achievements in mathematics, science (TIMSS), and reading

[70] https://www.openeducationeuropa.eu/
[71] https://ec.europa.eu/education/policy/strategic-framework/expert-groups/digital-skills-competences_en
[72] https://ec.europa.eu/education/policy/strategic-framework/expert-groups/modernisation-higher-education_en
[73] http://www.oecd.org/pisa/
[74] http://timssandpirls.bc.edu/about.html

(PIRLS). TIMMS testing was undertaken in 1995, 1999, 2003, 2007, 2011, and 2015, while PIRLS has been in 2001, 2006, 2011, and 2016.

While the international testing programmes provide good country-level comparability, and they have significant political influence at country level, the assessments are time-based (relating to a particular year, and with results usually not published until the following year), and it can be difficult to link the learning lessons to the delivery of equity, in real time, and at the individual level of the student.

The end result of existing approaches to monitoring has been a time lag, where policy decisions and system evaluations will be based on older, and potentially out-of-date, data. PISA, for example, takes place every three years, which meant that prior to December 2016 much research and policy was still referencing the results of PISA 2012, and even from December 2016 PISA results from 2015 will be referenced by policymakers as being 'current', and also more in the context of country's position in a ranking list (ECONOMIST, 2016a, OECD, 2016e).

Equity and efficiency are therefore critically important considerations for education systems, but have been difficult to monitor consistently (for example, in a timely manner) and also comprehensively (monitoring in detail at learner level in a way). Policy makers need to know how their education **system** performs against others in the world, and to identify what policies and practices they can consider to help improve it.[75] They also need to understand how their education system effectively delivers equity to all **learners**, irrespective of their needs.

That requires more than time-based samples of data. Susan Durston emphasises that equity monitoring must be fully institutionalised, and must be intersectoral (Durston, 2014). It must be **longitudinal, and capable of providing 'early warning signs'**. She cites the example of the UNICEF Education Parity Index as providing a range of indicators (although the indicators are very much time-based data[76]) which are at national level. Durston further recommends that an education system must 'listen' to data from all relevant sectors, and must provide **feedback loops** to parents, communities and policy. Durston warns that many equity needs cannot wait for annual surveys or for research projects to report – the needs are immediate and often severe (Durston, 2014). As UNESCO emphasises, it is essential to build a "*focused, evidence-based and dynamic monitoring and evaluation system for the education sector in order to adequately meet the demands generated by the new challenges*" (UNESCO, 2016a).

To explore those issues the following subsections review the predominant approaches to monitoring equity and efficiency, exploring emerging approaches using big data, and looking to the future potential for monitoring efficiency and equity in a holistic and individualised manner.

Monitoring equity and efficiency

Currently, there are not specific pan-EU levels indicators addressing overall equity and efficiency of education systems. There are limited examples at national level, and recent Eurydice report on structural indicators observes that the Quality Assurance Agency in Belgium (French and German speaking communities) does have an indicator for the HE level that is related to efficiency and equity, and which "*evaluates the processes and mechanisms in place within programmes to monitor student progress, including whether they successfully complete their studies*" (Eurydice, 2016e).

[75] For example, through the PISA Programme for School Improvement https://www.pisa4u.org/
[76] See page 28 of https://www.unicef.org/rosa/New_BeyondGender_09June_08.pdf

Equity

The OECD argues strongly that "*along with high overall student performance, the best education systems in the world also strive for equity in the performance of students of all socio-economic backgrounds and efficiency in public spending on education*" (Craw, 2017). Achieving **equity** is core to **UN Sustainable Development Goal** 4.5, where by 2030 the aim is to overcome gender disparities, ensure equality of access for all (irrespective of ethnicity, disability, family context, socio-economic background etc.) to education and vocational training (UNESCO, 2016c). However, UNESCO observe that while country-level comparisons are possible by "*sex, location and household wealth*", effective metrics are not yet available comparatively for "*disability, migration and displacement, language and ethnicity, citizenship status*" (UNESCO, 2016c).

Harmonising data across systems: For the EU, the building of evidence is complicated by the task for Eurostat of harmonising data across Member State statistical systems, meaning that the 'latest' data, such as gender breakdown of primary school teachers, were published in October 2016, but relate to 2014 (ESTAT, 2016d). Europe 2020 benchmarks focus on early school leaving and tertiary level attainment[77], while the ET 2020 benchmarks[78] include indicators that can relate to both equity and efficiency, although as of February 2016, the latest indicators related to data from 2015, further emphasising the often 'historical' nature of existing means of monitoring education systems. They include:

- Early school leavers (age 18-24); Early childhood education and care (age 4+); % of underachievers in reading (age 15); % of underachievers in maths (age 15); of underachievers in science (age 15); Public expenditure on education (as % of GDP); Expenditure per student in ISCED 1-2, 3-4, and 5-8 (€ PPS); Early school leavers (Native-born) and (Foreign-born).

For equity, a recent study noted that most EU states do not have strong monitoring systems that provide **comprehensive data about the progression and attainment of disadvantaged groups** across all types of education (Budginaitė et al., 2016). The heterogeneity of both education and social systems across countries was noted in a study of indicators spanning nine counties (Canada, China, Finland, France, Germany, Italy, Japan, United Kingdom, United States) and six data dimensions (Economic Equity; Social Stress; Support for Young Families; Support for Schools; Student Outcomes; System Outcomes) (HML, 2015).

The results of that study were less a ranked list of countries, and more a descriptive dashboard of their overall situation. For example, France was described as lagging behind the other countries on student outcomes and system outcomes, whereas the UK and USA were summarised as presenting the tension noted earlier in this section about balancing efficiency and equity: "*high levels of economic inequity and social stress, combined with commendable indicators in 3 areas: support for families, support for schools, and system outcomes*" (HML, 2015).

Nevertheless, the data needs in assessing equity are complex. The monitoring of equity at the level of an education system needs to be **strongly guided by the policy questions being asked of the data**. Students who are experiencing equity challenges in education exhibit a wide range of characteristics, and a student with a mental illness or physical disability, or one who has experienced abuse and who is withdrawn socially and educationally, can be from any social or family background.

[77] https://ec.europa.eu/info/strategy/european-semester/framework/europe-2020-strategy_en
[78] https://ec.europa.eu/education/resources/key-indicators and http://ec.europa.eu/eurostat/statistics-explained/index.php/Europe_2020_indicators_-_education

Monitoring equity can therefore be effective when it **monitors the individual characteristics and needs of students**. For example, Portugal has developed programmes which are individualised for students with special educational needs, and targeted support (for example in areas of education and health, evaluation) is provided within the regular school system (OECD, 2014b).

Dyson and colleagues (Dyson et al., 2010) further emphasise that monitoring equity must be **responsive to individual and local dynamics**. Since the reasons for inequity are multi-faceted, monitoring they involve community, family, and personal issues with which the education system must engage to deliver the education best suited for individual learners. They advise that the learning must develop a pathway that provides equitable career pathways, for example that learning opportunities need to offer opportunities within a local and regional context which would maximise the opportunities for school leavers to become local entrepreneurs, as well as local employees in companies (Dyson et al., 2010).

For students with a variety of equity challenges (disability, behaviour, family circumstances etc.) the associated data often are highly fragmented across multiple services. Agencies involved range from healthcare, social services, police and justice systems, to people such as carers. Interventions for students mostly involve physical meetings between the agencies, and each agency tends to collect its own data relating to the students, in their own format, which makes joining-up the data very difficult. **Fragmentation of information, and a lack of coordination and data sharing across services**, has led to serious abuses of children in schools and within society not being identified, even if the information was there but spread across agencies. Enquiries into the reasons behind service failures range from the Ryan Commission[79] in Ireland, to the ongoing and complex UK enquiry into historical cases of child abuse.[80] Equity therefore involves the safeguarding of children, as well as the provision of learning opportunities to all students.

The most **internationally authoritative, and comparable** (using PISA data), assessment of equity is provided by the OECD through country analyses reported in the *Education Policy Outlook* series. The variables used range across: the first age when selection is undertaken in schooling; the percentage of students performing high/low in mathematics; the variance in mathematics performance between and within schools as a % of the OECD average; the percentage of students who must repeat a grade at all school levels; the percentage variance in PISA mathematics performance explained by economic, social and cultural status (ESCS); the score differences between non-immigrant and immigrant students once an adjustment is made for their socio- economic status; and the differences in scores between males and (OECD, 2015a).

The PISA data is used in a dashboard to compare countries[81], although the data comparisons will depend on whether countries participated (or will participate) in one or more of the seven waves of assessment between 2000 and 2018.[82] PISA equity measures provide strong education system-level comparisons of what has happened. However, it is at the individual level where data can relate to individual students and their learning, and where other educational and social needs need to be overcome to deliver 'equity' to them.

PISA is determined not by building on country level data, but through undertaking **standardised tests**. The outcomes of the tests are regarded as providing insights into the level of equity in an education system, and do not link directly to the way in which an

[79] http://www.childabusecommission.ie/index.html
[80] https://www.iicsa.org.uk/
[81] http://www.oecd.org/edu/profiles.htm
[82] http://www.oecd.org/pisa/aboutpisa/pisa-participants.htm

education system assesses equity. Furthermore, the ways in which equity is provided to students through a variety of services (health, social, justice, as well as education) will have an impact on the efficiency of the education system.

At country levels, there are data ecosystems where data can be aggregated from the individual upwards, such as the "*Framework for Statistics on Learning and Education*" in **Canada**, where it is possible to aggregate data from the learner level to the institutional, administrative (what they term jurisdictional) and programme levels. The Framework includes data about ethnicities and disabilities (CMEC, 2010). However, as sophisticated as the system is, its focus is still strongly on the management of the educations system, on the allocation of resources at administrative or institutional levels.

While the Canadian Framework does link to data such as crime and social outcomes at the administrative level, it is not possible to disaggregate those data down to the learner level. Learners are therefore more the 'ground level' data providers for the Framework, rather than the Framework being the source of data intelligence that enables individually focused learning strategies for individuals that deliver them equity. That potential exists more in a big data approach, and this is developed later in this section.

Germany has a federal education system, with the 16 Länder having their own education policies, and coordination at the national level is undertaken through the Standing Conference of Ministers of Education and Cultural Affairs (KMK). There are comparative examinations undertaken by Länder, which enables comparability. Statistical data collection is undertaken by Statistical Offices at both national and Länder levels, as is reporting. In its review in 2014, OECD reported that policies at the system level were negatively effecting achieving equity, notably the tracking of students (a theme of another contribution to this study), selecting students academically, and requiring grade repetition where students did not attain the required level.

The 2014 OECD review noted that Germany had a highly diverse population, with immigration being a significant contributor to diversity. However, the PISA 2012 results had shown that "*students with an immigrant background scored 25 points less in mathematics than native students*", and only 13% of children aged under 3 were attending day-care facilities (OECD, 2014a). The importance of early childhood care is widely understood as being a fundamental building block for educational attainment, but in this non-compulsory yet formative area of education the data ecosystem will be highly variable, with limited educational data available for those children who are not participating in early childhood education and care (ECEC).

The OECD review for **Slovenia** focused on the challenge for effective use of resources, "*allocating them where they will have the greatest impact on equity and quality in education*" (OECD, 2016c). The policies enacted to deliver this have included new methods for financing upper secondary schools (moving to a funding regime per student and providing funding in blocks), teacher re-training. OECD observed that there is an accompanying need for "*improved information on the number of students and the real needs of the system*" (OECD, 2016c)

In **Hungary**, an equity requirement introduced in 2015 obliges schools to provide evidence and analysis relating to student attainment, their social characteristics and their family context. A particular equity challenge for Hungary exists with the Roma, who consistently underperform as an ethnic group, having low levels of pre-school participation, high levels of early school leavings, and often have challenging family contexts (OECD, 2015b).

Efficiency

UNESCO has an **efficiency methodology** for monitoring national education accounts (NEA), particularly for countries that are building education systems. Data domains and analytical approaches are provided, with a focus more on the system level monitoring of

finance through data aggregation from conventional administrative sources, but also identifying data relating to the student characteristics (socio-economic characteristics etc.). Assessing the costs per class can help assess efficiency, as can average teacher salary costs. Participating in international assessments such as PISA can provide independent external evaluation of standards at the system level, and national tests and exams help to understand efficiency and effectiveness at the institutional levels upwards (UNESCO, 2016b).

Efficiency can be monitored formally through a **school inspection process**. New Zealand does this through a set of indicators[83], with an online facility to see the results in school profiles.[84] The 55 indicators are a long and complex list of themes, ranging across annual expenditure per student, "*impact of education on income*", mathematics and literacy achievement, the provision of services for early childhood education, to the level of youth suicide. The profiles are dependent on the date when the school was last inspected, so the data appears only at institutional level and at the time of a report (for example, a Christchurch school with data for 2015[85]). The school inspection process therefore provides standardised data for schools, more in the form of a dashboard per school rather than in a 'league list'

The UK Ofsted[86] also monitors schools through a standardised framework, looking at criteria such as how they **performance manage** teachers, how that performance is linked to career and salary progression, and how it is used to promote excellent practice among teachers. Evidence gathered includes the professional development undertaken and how that contributes to better teaching, how the head-teacher interacts with governors, leadership teams, and teachers and other staff. It looks at the systems schools put in place to track and monitor the impact of teaching, how evaluation is undertaken, what consultation methods are used, and records and data are kept.

Efficiency monitoring can also be undertaken mainly through a **top-down process** (setting standards through a national curriculum in addition to monitoring standards through independent inspection regimes) and through **bottom-up aggregation of data** through standardised administrative returns. A school is then assessed more on the basis of its aggregate performance (as an institution) rather than on an individual basis (teachers, support staff, students).

There are risks that those within education systems start to **adjust behaviours to maximise performance** against the official monitoring frameworks. They may focus curricula on the 'core' targets, stopping the teaching of other less important in the performance metrics. A well-resourced and highly selective private school in effect starts with an expectation that all students will meet examination targets. A local state school, taking students from its geographical catchment, has less control over intake quality, and more challenges in value-adding each student up to the target levels.

When efficiency is measured against national targets (for example, x% of students in a school at a certain level achieving y number of examination grades above a certain level) the resulting metrics do not adjust for local differences, such as levels of exclusion, poverty, family situation, medical and behavioural issues. This in itself creates an equity issue because it fixates performance on absolute levels, and not on the extent to which students have experienced **value-adding in a school**. In Scotland, there is clear attention to value-adding through overcoming the gaps in educational attainment experienced by disadvantaged children. Interventions will be evaluated through the

[83] https://www.educationcounts.govt.nz/statistics/indicators
[84] https://www.educationcounts.govt.nz/find-school
[85] http://www.ero.govt.nz/review-reports/pegasus-bay-school-30-11-2015/
[86] https://schoolgovernors.thekeysupport.com/school-improvement-and-strategy/inspection-evaluation/inspection-framework/report-to-governors-on-teacher-appraisal/

analysis of data such as attendance, school inspections, numbers of exclusions, self-evaluations of schools and local government structures (SCOTLAND, 2016).

Yet, education systems place importance on achieving **social mobility**, since with social mobility comes labour market mobility. It is a complex issue, relating not just to education, and the current UK situation reports an entrenched problem of inequity ranging from the early years through to tertiary education, noting that "*only 1 in 8 children from low-income backgrounds is likely to become a high-income earner as an adult*" (SMC, 2016). For the EU, the monitoring at the system level of inclusive education has been a political priority, but it still remains unclear what themes the monitoring will assess, and how they will be measured (Watkins and Ebersold, 2016).

Efficiency is influenced strongly by teachers, who as a workforce may be strongly unionised, sometimes with low wages (teaching not being an attractive profession), or lacking the teaching and learning infrastructure to help them become more efficient. Where labour (teachers) is organised at a system level there can be time lags in implementing modernisation, for example as is the case in Latin and South America countries such as Chile, Ecuador, Mexico, and Peru have passed legislation to link teacher remuneration to performance, but "*none has yet had the courage to implement a rigorous evaluation system under which teachers who fail are ejected from the profession*" (ECONOMIST, 2014).

A Latin America study by the World Bank found that in spite of good resources in many schools, the teachers spent less than 65% of their time 'teaching' compared to the US benchmark which is 85%. Problems in achieving improvement included poor recruitment and training methods, and resistance to change from powerful unions: "*A sine qua non is national testing of students and the publication of schools' results*" (ECONOMIST, 2014). Strong resistance to organisational change was evident in Mexico, which had to overcome a practice where there had been a process whereby the teacher unions were in charge of recruitment (ECONOMIST, 2016c).

Effectiveness has been linked to the **quality of teacher training**. A a meta-analysis of 65,000 research papers found that "*teacher expertise … what a teacher did in the classroom*" was the single most important factor in delivering quality learning to learners (ECONOMIST, 2016d). Other research also confirmed that selectivity of students does not guarantee better learning outcomes, whether a school is state-funded or run by for-profit organisations (Boeskens, 2016).

Education systems have looked at various mechanisms to **'empower' efficiency**, by removing from local or regional government any administrative control over schools, and delegating powers to the local level. The theory behind this (the theory of action) is that more autonomy, and less governmental interference, would enable more rapid innovation in teaching and learning, leading to more efficiency when delivering high quality and equitable outcomes.

One approach has been to move away from national salary scales, to more competitive appointment systems. There have been attempts to raise the efficiency and effectiveness of schools that are in challenging areas, or to overcome problems in the supply chain of STEM expert teachers, for example by providing more attractive contracts to those willing to teach in such schools (Kirby and Cullinane, 2017). At the organisational level, there have been initiatives such as USA 'Charter Schools'[87] (mostly physical schools, but also virtual 'cyber' schools are allowed), which are privately-run (both non-profit and for-profit) schools that are still funded by the public sector, and are subject to the same standards as other schools.

[87] https://www2.ed.gov/about/offices/list/oii/csp/index.html?exp=7

However, in a review of US Charter Schools over an eight-year period, Welch found that by insisting that the Charter Schools are accountable against the same performance metrics as other schools, it ended up with Charter Schools 'behaving' very similarly to schools funded in the public sector, with Welsh warning that "*reform models often get bogged down in measurements of fidelity rather than efficacy*" (Welch, 2011). This observation presents a monitoring dilemma. Consistent and comparable monitoring can be achieved if all types of school respond to the same monitoring process.

If a radically new school approach is taken, should it define its own monitoring metrics, and focus on value-adding (efficacy) as well as activity and performance levels? Part of the problem in responding to the question is not so much that one approach excludes the other in principle, but that the data mostly do not currently exist to enable both, because data is provided at too aggregated a level. This is where the potential of big data exists, to monitor seamlessly from the individual level (student, teacher, manager) to the system level, and potential approaches are introduced in the next section.

In educational data, metrics such as student attendance, grades and retention (at secondary and tertiary levels) can be an indication of the engagement, motivation and talent of a student, but it can just as equally indicate poverty, family violence or absence, geographical isolation, and a number of other factors linked to socio-economic disadvantage. As Har Carmel writes, a monitoring system that uses big data needs to be "neutral" in the algorithms and the data domains used by analytics, and not to "*rely on biased data that reflect social inequality and plausibly reinforce present structural inequities and contribute to a problem of cumulative disadvantage*" (Har Carmel, 2016).

Developing big data monitoring

Research by McKinsey suggests that improving the utilisation of education data, and in particular using open data, could contribute between $900 billion and $1.2 trillion into the global economy each year (Fassbender and Giambrone, 2015). With such a level of potential benefits there are clear motivations for policy makers to monitor their education system more efficiently and effectively, and to deliver equity more directly to those who need interventions and support.

However, pragmatism is important, and finite resources, combined with political realities, often drive decisions rather than the need to deliver optimal outcomes for learners. EU democracies have governments that are elected only for a limited number of years, and the time that is taken for significant organisational change and the design and competitive procurement of a new data ecosystem, is often longer that the period a government is in power before the next election.

Consequently, development of new big data monitoring approaches is conditioned by three drivers: first, where a government identifies an opportunity to use the existing monitoring system more effectively by integrating and interoperating across existing data; and, where circumstances are opportune for a government to take a radical decision to build a completely new big data system.

Developing information integration through interoperability

For the foreseeable future, many education systems will continue to rely on their existing data sources and data models, but may be able to monitor equity and efficiency more effectively through widespread data integration and/or data sharing. This approach used technology tools to build links across the data and organisational silos noted earlier in this section.

Making better use of existing data sources can be cost-beneficial. Data can be time-consuming and expensive to gather, generating administrative burden on data providers, which can lead to survey fatigue. Existing data sources may capture the same data more than once, and teachers cite data input and management as a major factor in heavy

workloads, in terms of the time spent managing student records, and compiling data for submission to national collections and evaluation schemes (ECONOMIST, 2016g). In the UK, for example, a stocktake of data collection requirements for higher education institutions found that there were 525 different separate collections and 93 different organisations collecting the data.[88] Making better use of what is available can therefore increase efficiency.

To link data better at the system level there need to be developed technical mechanisms that link together disparate IT systems (interoperability), to make the different data ecosystems 'communicate' with each other (using standard terms and meaning – ontology), and for the records relating to individual people (students, staff etc.) to be clearly identified as being related to them (identity management – identity numbers) to ensure that the data that are linked to individuals definitively relate to them.

Cantini and colleagues have examined the potential for big data in monitoring the Italian education system, noting the legacy of diverse and heterogeneous data sources that, if fully integrated, could facilitate a better monitoring of the education system. To achieve that, however, requires attention to semantic interoperability across diverse data, and the generation of a common ontology for the data structures (Cantini et al., 2016). They note the existence of the European Interoperability Framework, which helps to promote cross-border service development through the interoperability of public services, and linking national interoperability frameworks (Commission, 2010).

There are explorations in building big data resources through linking existing, but separate data series. The World Bank, in upgrading its EdStats[89] site to provide what it calls a 'big data' approach, making available "*education indicators (enriched with learning data) on one platform*" (Abdul-Hamid, 2014).

The Digital Strategy for the UK is undertaking a multi-sector approach, through the construction of a "Data Exchange" where common and open standards will facilitate interoperability across different IT and data systems in the education sector, also minimising the administrative load on those providing the data in the first place. The Exchange will be developed partnership with schools and software suppliers (DCMS, 2017). To further build insight into data linking it announced in March 2014 that secure access will be made available (through the Virtual Microdata Laboratory service[90]) to test the potential of big microdata in providing more effective insights into the education system (DCMS, 2017).

In 2016 the UK published the first big data investigation from its longitudinal educational outcomes dataset, which is linking education data with information from a number of government departments, including tax data (DfE, 2016). The aim is to provide an accurate picture of graduate destinations in the longer-term, and particular graduate earnings. Records from the Higher Education Statistics Agency (HESA) student records are matched to Department for Work and Pension (DWP) Customer Information System (CIS). The matching suffers from a lack of a single identity number in the UK[91], requiring an initial linking of records through an algorithm rather that a unique ID.

An example of where better linking of existing data has worked for many years at the country level, is the Crossroads Bank for Social Security in Belgium. While this example is not explicitly relating to the education systems, it shows that effectively linking existing data in a secure environment, can work: but only if data can be collected, aggregated,

[88] https://www.hediip.ac.uk/
[89] http://datatopics.worldbank.org/education/
[90] https://www.ons.gov.uk/aboutus/whatwedo/paidservices/virtualmicrodatalaboratoryvml
[91] Proposals for an ID card in the UK failed in 2004 after privacy concerns and public hostility. However, where data linking is to be undertaken, having a single and unique ID number across administrative systems significantly helps the matching process.

integrated, and processed more efficiently than before. This requires work on interoperability, ontologies, analytics, data protection and IT security.

The Crossroads Bank for Social Security[92] (CBSS) does for social security what could be undertaken for the linking of education system data sources: it works seamlessly across multiple sources of data to provide individual level real-time monitoring of needs for support from the social security system. The social security system involves over 3,000 institutions, with different data collection regimes, different computer systems, and nearly 30 years ago, a review of the system noted that the processing of data was slow, inefficient, did not lead to 'customer focused' service delivery. Citizens were often asked for the same information many times. However, it was also understood that to make all of the institutions move to a common data regime would have required massive organisational change and system re-engineering.

Instead, the approach taken was to create an independent and trusted intermediary who could work across all the administrative data systems through a process of interoperability. The system would not need to create a new data regime, nor would it need to combine all the data into a single massive IT system. The IT system of the Crossroads bank instead knows about the data models of all the institutions. When a query comes in about the change in personal circumstances of one individual (and the identity card is the data link across all the institutions, and a common identity mechanism is essential for such a process), messages are passed to the systems of the institutions.

The bank is the agency that exchanges messages between social security institutions, and an answer is constructed, for example about whether that individual should receive a new benefit. If they are eligible for the benefit the system automatically allocates it and send a message/letter to the beneficiary.

The equity gains have been huge, since the beneficiaries do not even need to apply for many services, and are told proactively when they are eligible to receive one, with significantly more precision than before (avoiding the contradictions that previously existed when being provided with one service led to another service being removed or reduced). The efficiency gain was significant, with the elimination of over 220 paper forms that were filled in manually before the bank was created, and dramatic speed-up of decision-making "*in 2016 1,109,577,113 concrete electronic data exchanges took place with a response time for the online messages lower than 4 seconds in 99.27 % of the cases*". [93]

Developing big data systems

Building a completely new big data focused monitoring system can require significant organisations re-engineering. However, two examples from Portugal and Estonia show that it can be achieved.

An example of a country making a transition from decentralised data to national big data is the system being developed in Portugal. It has been developed at the request of the Troika following the economic crisis (Evaristo, 2014).

The original Portuguese education monitoring system was established in 2005, when there were not enough resources to establish a centralised database to which all schools were linked. Instead, a decentralised approach was implemented, with schools storing data locally and the extracting and reporting data from their local systems. Data were aggregated yearly at the system level. However, comparability across years was less

[92] https://www.ksz-bcss.fgov.be/en
[93] https://www.ksz-bcss.fgov.be/en

robust, because there was no single student identifier. As a result nearly 20% of records were not aggregated when data were combined over different.

Responding to the request from the Troika, and knowing that developing a completely new monitoring system would take time, Portugal first developed an interoperability solution, integrating existing databases (similar in approach to the Crossroads Bank above) and linking to the Finance Ministry database to track financial activities across schools. Dashboards were created for information domains: schools, their students, teachers, and other staff, examination results, activities relating to social support measures, and special needs education. This allows better aggregation from schools to the national level, with the primary use being at the Ministry level for managing the school system. However, while data is integrated and aggregated more efficiently, the overall effectiveness of the process is not dramatically increased since the frequency of data availability was not improved.

The significant big data development is the *SIGA @ Portal das Escolas* (Schools Portal), which is a new centralised system building on information relating to individual students, including data relating to family situations. Schools will be connected to SIGA, and the system will have data validation checks to build consistency, and all data will be coded to the electronic citizen card held by each citizen.[94] This will increase efficiency of managing the education system, and enable better equity through the availability of individual student data, as well as providing a consistent student data record as they move between schools. More importantly, data will be available in real time.

Data domains will range across biographical information about students, family situation (socio-economic), data about classes (composition, management), attendance, examinations and attainment. And, the system will have an alert system to inform schools and parents about important information, or the need to take action. The management of the education system will become significantly more efficient by having integrated and real-time evidence for the education life-cycle of each student, so that the Ministry of Education "*will have a tool to monitor early school leaving by implementing early warning mechanisms that can be managed either at central or school level*" (Evaristo, 2014).

A second example focuses on how a country can 'leap over' the legacy of outdated systems to become a world leader in big data application for monitoring its education system. As it emerged from Communism, Estonia took the decision to look ahead at the development of eGovernment services, and to build a fully-functioning information society, where the national identity card would become the basis of all data transactions, not just in the public sector, but also the private sector (for example banking). Being sensitive to a history of pervasive state surveillance of citizens, the Government of Estonia decided that the adoption of the identity card should be voluntary, and that take-up would occur through a rich variety of integrated services through IT systems that citizens trusted – the trust was reinforced when Estonia resisted one of the first cyberwarfare attacks on its systems in April 2007 (Ruus, 2008).

In Estonia, a big data approach has been built strategically as the country has become a leading developer of integrated digital solutions – eEstonia.[95] Equity is addressed through a long period of comprehensive education that minimises grade repetition. Statistics are collected through the Estonian Education Data System[96] which provides integrated access to data at student, teacher, and institutional level, and which has the ability to track student learning pathways. It embeds information from other institutional

[94] http://www.gemalto.com/govt/customer-cases/portugal-id From 2014 a single card replaced five documents: cards relating to civil identification, taxation, voting, social security, and healthcare.
[95] https://e-estonia.com/
[96] https://www.eesti.ee/eng/services/citizen/haridus_ja_teadus/isikukaart_eesti_ee_portaali

sources such as health[97], banks (Estonia has a very mature eGovernment system[98]), migration, and citizenship.

The Estonia eSchool[99] service was introduced in 2002, and as of early 2017 was used by 85% of schools who were teaching 95% of all students. eSchool provides a totally integrated resource, from the individual (student) to the country level, where:

> "*Parents can see their children's homework assignments, grades, attendance information and teacher's notes, as well as communicate directly with teachers via the system. Students can read their own grades and keep track of what homework has been assigned each day. They also have an option to save their best work in their own, personal e-portfolios. District administrators have access the latest statistical reports on demand, making it easy to consolidate data across the district's schools*".[100]

Underpinning trust and confidence in the system are the powerful security and privacy protections applied to the national identity card[101], with very strong system security and cyber-defences.[102] There is very strong access control to data (ensuring that only those authorised to use data can access them), and crucially there is robust transparency, where "*citizens have rights to access and inspect data held about them; transparency breeds trust, over time*" (Brett, 2015).

The Estonian Education Data System is used for monitoring purposes by the National Statistical Office, and the Ministry of Education. Aggregated indicators are available openly (an open data culture) on the HaridusSilm[103] ("Education Eye") website, providing all stakeholders with information about the performance of the education system (OECD, 2016b).

Challenges, policy actions and future trends

Challenges

In the conventional organisational environment of education systems location (a physical place where education is provided) remains a dominant force, and hence data is organised within a locational perspective. A school (physically located, and which recruits its students mostly from its local geographical 'catchment') is part of a governance structure (local or regional government, federal structure) which is part of a national system. Information filters upwards from location, and the most important 'atomic' data element (the student) is treated as being part of an aggregate structure (the school up to the system).

Furthermore, location also influences efficiency and equity. Where league lists are produced of school 'quality', the property purchase and rental prices can increase because parents want to move within the catchment area of a 'good' school. Data gathered by the UK Department for Education shows that the average house price in July 2016 was £233,000, but in areas where a school is inspected by OFSTED and graded excellent, the average house price "*near the 10% best-performing primary schools are 8.0% higher than in the surrounding area. Near the 10% best-performing non-selective secondary schools, house prices are 6.8% higher*" (DfE, 2017b). And, this is the premium for house purchases, and those families who are renting or are in public sector housing have significantly less ability to move to those school catchments.

[97] A fully electronic health record https://e-estonia.com/component/electronic-health-record/
[98] https://e-estonia.com/components/
[99] https://e-estonia.com/component/e-school/
[100] https://e-estonia.com/component/e-school/
[101] The card is used by 94% of all residents https://e-estonia.com/component/electronic-id-card/
[102] https://e-estonia.com/the-story/digital-society/cyber-security/
[103] http://www.haridussilm.ee/

That can distort equity, because those who can afford to move will gain preferential access to excellent schooling for their children. The more detailed, timely, and rich the data sources become, then the data users will find new ways of interpreting and using resulting indicators. Increased access to data, and a sense of entitlement to good teaching, can lead to contested positions, for example where parents cannot 'see' the monitoring information about their children other than when it is mediated by schools, so information is often contested rather than shared (Burns, 2016b). However, this is not a problem with the data 'per se' but with the ways in which results are used and interpreted, but in an era of 'fake news' the potential for data misuse increases.

Equity is also challenged by the difficulties of persuading excellent teachers to go to challenging and low-performing schools. It is quite natural for a high-performing teacher to focus on being employed in an excellent school. Consequently, there have been initiatives at recruiting excellent teachers to low performing schools (for example through salary incentives), supporting them with strong school governance, and using strategies to retain them working in the school (COMMONS, 2016). These are all place-based strategies and policies, for example aiming to overcome equity problems resulting from schools with a majority of students from disadvantaged or minority backgrounds. Since the students are fixed in location, the approaches have been towards encouraging the teaching and management staff to be mobile. Learning platforms and big data approaches can work to overcome such problems through blended learning approaches.

However, to create high-performing, comprehensive analytics systems, the systems first need to be 'trained', where historical data is fed into the system to teach the analytics what bits of data is significant, and to identify trends in the data from high and low performing institutions and students. As an example, if a school analytics system sees that students who are frequently late at Year 3 level generally receive lower test scores at Year 6, it might start flagging those students as low-performers at an early stage. At a higher level, a monitoring system might penalise a school which has a student body that on the whole performs lower than averages, despite the fact that the school is doing better on equity measures.

A study for the Joint Research Centre supports caution, observing that the black-box nature of most analytical algorithms risks generating results that "*can work against equality and equity*" (Ferguson et al., 2016), and a US study warns that whatever algorithms are applied in the context of assessment, the assessment mechanisms must also be equitable for all learners: "*our ambitions to capture learning have often outpaced our abilities to design effective assessment tasks*" (Thille et al., 2014).

Policy Options

In the immediate future, the EU Member States will display an uneven picture of data and big data usage in monitoring their education systems for equity and efficiency purposes. The monitoring landscape will change, with new leaders (such as Estonia and Portugal) having modernised their monitoring systems, and benefitting from a fully integrated big data landscape from student to country levels. As Estonia shows, being able to aggregate data from the individual level to the system level is built into the big data system. This should enable them to develop and implement new and innovative individualised teaching and learning for individual students efficiently and equitably.

It will present new challenges for the EU, since the existing monitoring models are based on data systems that are sample and time-based and pre-big data. However, there will be significant roles for the EU in developing richer monitoring while supporting Member States whose monitoring 'maturity' lags behind the leaders. There have been significant areas in which the EU has already been active:

- Identifying and setting common information standards, and building a stronger European evidence base for education policy making (Eurydice, 2017b);

- Building robust and meaningful indicators and metrics for Member States to use in benchmarking their education activities[104];
- Facilitating innovation and development in education technologies, for example through Erasmus+ Strategic Partnerships[105];
- Supporting Member States to maximise their investment in educational technologies and methods through activities under the Open Method of Coordination, for example through the ET 2020 Working Groups on Schools, and on Digital Skills and Competences[106].

The EU has been clearly aware of the potential of big data technologies and in 2014 outlined a new strategy in big data, designed to support and accelerate the transition towards a data-driven economy in Europe (Commission, 2014a). This strategy incorporates a number of important goals, noting that the EU should:

- *"Support "lighthouse" data initiatives capable of improving competitiveness, quality of public services and citizen's life;*
- *Develop its enabling technologies, underlying infrastructures and skills;*
- *Extensively share, use and develop its public data resources and research data infrastructure"*. (Commission, 2014d)

For example, the European Data Protection Advisor has promoted a "*Digital Clearing House to bring together, for the first time, agencies from competition, consumer and data protection areas who are willing to share information and discuss how best to enforce rules in the interests of the individual*" (EDPS, 2017).

However, these important developments are not fully joined up in a way that would provide a clear focus on monitoring equity and efficiency of education systems. While the big data strategy makes note of sectors such as health, transportation and logistics and agriculture and food supply, there is not a coordinated approach in education as to how this sector could benefit from the increase in big data in monitoring education systems.

In the current generation of ET 2020 working groups; two focus on equipping the education sector for the future, with one group looking at the modernisation of higher education and the other investigating digital skills and competencies. A third looks at promoting common values such as tolerance and non-discrimination, which is relevant for equity. Following the interim evaluation of ET 2020, the EU adopted six new priorities in education for 2016-2020 (Council, 2015b), two of which highlight equity and efficiency.

These are strong policies, aiming to help the education systems in Europe continue to modernise and cope with current and future challenges, but the policies and strategies are not coherent on how big data technologies can support the sector in its progress towards these aims.

In terms of supporting the building of monitoring approaches, the EU is in a good position to lead on setting common standards for data users and developers, as well as supporting semantic interoperability and ontology developments noted earlier in this section. Common education data standards would outline the core data that can (and cannot, for ethical reasons) be collected, provide guidance on the collection of other data, and detail appropriate methods for collection, storage and analysis. This could help to ensure that everyone is operating to a common standard in their use of education data

[104] https://ec.europa.eu/education/resources/key-indicators
[105] http://ec.europa.eu/programmes/erasmus-plus/opportunities-for-organisations/innovation-good-practices/strategic-partnerships_en
[106] https://ec.europa.eu/education/policy/strategic-framework/expert-groups_en

analytics, which in turn allows for the data to be confidently used in monitoring systems. If done well, it should also help prevent against hidden biases and entrenched discrimination finding its way into Europe's educational monitoring.

EU action in the areas of big data and education systems over the next 10 years could focus on sharing the good practice seen in Portugal and Estonia, and establishing cross-border standards, ontologies, and frameworks for anonymization, data sharing and interchange so that the big data systems conform clearly to the General Data Protection Regulation (detailed in the privacy section). It will also need to cope with what is already a differentiated landscape of national monitoring systems: big data world leader with Estonia, emerging integrated big data with Portugal, interoperable systems, and legacy systems.

There would need to be political agreement across Member States about what should be measured and analysed at education system level, rather than taking the data produced by the learning systems. It will be important to work with the Edtech industry, to improve the functionality of their systems and to build interoperability across data domains (for example developing core vocabularies), Accompanying that could be OMC actions to share expertise, experience and good practice.

The future

A big data approach could first (in the next 10 years) mandate 'privacy by design' (the privacy section develops this), and develop the interoperability frameworks, while supporting those Member States that wish to build on the systems already in place in countries such as Portugal and Estonia. The widespread adoption of an 'atomic' level of student monitoring would not just look at their educational performance through learning platforms, but would also look at issues of value adding, and flag learning issues, relating them to data from other relevant data systems and assess whether the learning issues are purely educational, or are influenced by other social or external issues.

Big data monitoring systems are not a 'possible' development – they are already in place and are being implemented – Estonia and Portugal clearly show that. Over the next 30 years that should lead to some potentially system-changing ways of delivering equity.

At present the equity needs of students are assessed at the 'ground level' through the interaction of agencies which usually have their own data cultures. Data is seldom shared in real time, interventions for students are planned and implemented slowly, and the funding for the students does not usually follow directly.

Big data monitoring of students and of classes and teachers could enable real-time allocation of resources. Instead of school funding being passed down from higher administrative levels through block grants, and reviewed yearly, funding for student needs could follow a big data decision that the needs are real and the resources are required. Crossroads Bank for Social Security shows that this approach can work in actuality even with existing data systems, and Estonia shows that data from the student can be linked to the system level in real time. This could deliver significant efficiency gains in delivering equity.

It could also provide students at school with blended learning, where specific needs are met through ICT innovations as well as through place-based learing. With the rapid developments in artificial intelligence, 3-D printing, and robotics, learning could become significantly more equitable through technology efficiency gains. Facilities such as automated translation (overcoming an English language dominance of content), culturally relevant content and pedagogy, voice recognition, and even gesture and brain scanning technologies (WIRED, 2017) could help to integrate learners who often are marginalised from mainstream location-based education facilities. Learning could be progressively less place-based in physical schools. Such developments would be strongly supported by a continuation of Commission initiatives in areas of inclusion or technology 'design for all'.

Location-based dominance of learning would decline, and excellent teachers could be linked to challenging schools regardless of their location.

This could have dramatic impacts on what we see as a 'school'. A school as a physical body is a legacy of friction of mobility. Learning has mostly (with the exception of institutions such as private and/or residential schools) been constructed on the basis of a 'catchment' or a neighbourhood.

Nevertheless, locality is a strong basis for identity, and there is a strong socialisation role for schools. Consequently, this is not a call to abolish place-based learning and go totally into an individually-monitored online education system. But, where all students and all schools use (as with Estonia) the same monitoring system, then students with special needs could be taught partially as special virtual groups by teachers whose expertise most matches their needs.

The potential threats of pervasive data need clear acknowledgement. As other contributions in this study advise, the pervasive, geographically tagged, and rapidly updated individual big data can be a source both of benefit and of more hostile surveillance. Teachers may have the same fears as workers in other areas of the labour market that are experiencing pervasive surveillance. Such concerns have led to the European Parliament reaffirming fundamental rights in the context of big data, warning that "*the trust of citizens in digital services can be seriously undermined by government mass surveillance activities and the unwarranted accessing of commercial and other personal data by law enforcement authorities*" (EP, 2017).

If all the data is integrated in real-time from the learner and the teacher upward, could 'algorithmic management' (O'Connor, 2016) lead to the abolition of head-teachers and school managers? Can those students with more challenging inclusion issues be better served by a portfolio of online and offline services mediated by multiple learning advisors? How will we perceive the progressive merging of the administrative and personal data environments? If education was privatised that this becomes less a public-private issue?

This would have implications for governance, since a head teacher may no longer be in charge of a school, but of a set of school services. The management of the physical infrastructure of a school could be put out to competitive tender (efficiency), allowing the education staff to focus on delivering quality education (equity), although contractual relationships would need to be carefully considered: in Scotland (UK), private finance was used to build and run schools but the debt for a large number of schools was sold to offshore investment funds (BBC, 2016). The 'dimensions of unintended consequences' will need careful consideration as big data monitoring of education systems develops further.

Assessment

Introduction and Context

This section examines the potential of big data is the assessment process to enhance the quality and monitoring of education systems. It sets assessment within the broader context of new teaching and learning systems, and in the ways that assessment is both undertaken (for example, how students can perform new types of assessment and have new behaviours towards assessment), to the ways in which assessment is 'assessed', and the ways in which the resulting assessment data or metrics can be analysed and used. It further considers how assessment processes and outcomes are 'owned', particularly where the assessment process goes beyond administrative or national borders.

The section initially considers how assessment issues emerge through the lens of experience at the edX[107] platform, which has higher education online offerings from over 100 higher education institutions worldwide. It then examines the issues at the more complex and heterogeneous levels of schools and school systems. Finally, it sets out some policy challenges at the European level.

Educational technology is evolving rapidly, and with that change, educational data is changing. A decade ago, educational data sets consisted of fairly simple data, such as student submissions of multiple choice answers and numeric answers (Koedinger et al., 2010), or data sets about overall student performance in school districts. Today, educational data cover a much broader range of activities, such as on-line student social interactions, including text, audio, and video data, and fine-grained interactions, while solving authentic assessment problems.

Broadly speaking, over the past few decades, teaching-and-learning has moved student learning to a more cognitively active process, generally with rapid feedback, paths for remediation, mastery learning, and the ability for students to self-pace. Such pedagogies, whether technology-enabled or not, had resulted in an approximate doubling of learning gains even back in the 1990s (Hake, 1998).

However, they are impractical to apply in educational settings without supporting technology. One-to-one tutoring shows a one-sigma or two-sigma gain over a traditional lecture format, depending on the study (Bloom, 1984), but requires a tutor per student. Other techniques, such as peer instruction (Mazur, 2009) show large learning gains, but require complete retraining of instructors. This is difficult, especially in a university setting where courses are taught by subject matter experts, but requires significant retraining even of professional educators at K-12 levels as well.

Learning at scale is defined as having substantial course resources (increasingly, entire courses) shared by thousands of students. Digital at-scale learning systems allow broader-based application of evidence-based techniques and show similarly large learning gains. In early work, this was constrained to intelligent tutoring systems in mathematics and physics, with relatively high content development costs, where such systems achieved gains approaching those of human tutoring (VanLehn, 2011). However, over the past half-decade, we have a growing number of techniques which allow the use of technology to enable the economic application of such principles at manageable costs (Mitros et al., 2013).

It is difficult to overstate the potential positive societal impact. If we can improve learning by a mere 30% (which is not an uncommon result in deployments of such systems) high school students would graduate with knowledge bases equivalent to our college graduates. In evaluations of the edX platform, results showed significant learning

[107] https://www.edx.org/

gain in on-campus use. In a blended learning trial at San Jose State University, course completion rose from 59% to 91% (Ghadiri). Even in pure on-line settings – with no human support – gains were higher than those of traditional in-person courses, although not as high as blended courses (Colvin et al., 2014).

While historically such systems focused on relatively narrow domains (such as simply concepts in mathematics and physics education), the richness of such systems has grown at an astounding pace. The following examples illustrate the types and scale of data generated:

- edX has over 1000 courses from 100 institutions with 10 million enrolled students, from every country in the world. The educational data captures minute interaction, such as each time a student views a video, a page of a textbook, or submits a problem or assessment. It does not capture individual keystrokes or mouse motions, but some platforms do. The edX dataset is several terabytes in size, with a few gigabytes per course;
- RichReview is a system where students can discuss and annotate documents. For example, one student might upload a Supreme Court decision. A group of students can then discuss that document with either text annotations, or by adding a voiceover. During the voiceover, students can write on the passage (with a stylus), point to specific passages (again, with a stylus), or highlight text. Such annotations can form a rich semi-asynchronous, interactive discussion about the document (Yoon and Mitros, 2015);
- Piazza is a course discussion system widely used in classrooms. In addition to posting and responding to posts, students can post wiki-style answers which other students edit;
- Google Docs is widely used for collaborative groupwork. Google Docs provides APIs which allow tracking of student contributions. Learning analytics built around these APIs allow instructors to visualize who did what and when (McNely et al., 2012);
- Video conferencing is increasingly used in educational settings, both for distance learning, and to connect students across cultures, disciplines, and campuses. Such data is generally not stored, but costs reached a breakpoint to where this is becoming feasible. Studies suggest that such videos could be mined to analyse turn-taking dynamics, affect, non-linguistic social signals, and other properties of student interactions to give helpful formative feedback on the development of soft skills (Pentland, 2005).

What is perhaps most astounding, is how personal and networked the learning data have become. Students use online submissions systems for essays about their personal lives, discussion systems to argue politics, and group collaborations systems to work on projects with real-world impact. Such discussions may devolve into swearing, bullying, or harassment. Students may express socially inappropriate views on political subjects. In other words, educational technology has moved from a space where such data is relatively safe, to one where it contains highly private information which could be damaging to students' future careers, family lives, and psychological well-being. Compare this to traditional educational datasets consisting of correct and incorrect numerical answers and timestamps (which is still often presumed in discussions of technology and policy).

Historically, educational policy was driven in part by educational research. Such research looked at data sets for large populations of relatively coarse data (demographic, school enrolment, course grades, and similar high-level metrics), and drew high-level conclusions, for example, comparing the efficacy of different models of charter schools. This was due to the lack of other types of data. Now, educational data are minute in

granularity, showing click-by-click and keystroke-by-keystroke interactions, often giving insights into students' problem-solving processes, group skills, creativity, and other higher-level skills (as well as in-depth observations on what happens in classrooms).

The range of such data is broad, covering the incredibly diverse tasks we listed above from virtually all of a student's courses and learning activities. Such data is also longitudinal, covering students' experience from preschool through adult learning. This has the potential to be integrated into the assessment process, so that student performance is assessed continuously, rather than only at certain fixed milestones such as essays, lab reports, and exams, revolutionizing education systems, education researcher, and education policy.

Unfortunately, we are still a long way from achieving such a goal. While quantitative decision making is well-established in business, entertainment, and engineering, there are substantial organisational, policy, and human capital roadblocks to similar advances in education.

For example, while new forms of rich assessment data are being gathered, they are primarily gathered by for-profit corporations. Educational data are considered proprietary, and while the bill for developing such technology comes indirectly from taxpayers, few corporations share it with governments, researchers, or the students to whom such data pertains. Students and teachers merely have access to aggregate results. Since such data are divided among hundreds of educational technology corporations, there is no easy way to combine or correlate such data.

Largely due to the success of technology in improving education, it has been widely adopted, and by 2008, there was approximately one computer for every three students in K-12 schools in the United States (NCES, 2014). Even that is an understated statistic, since it misses the similarly high penetration of student-owned digital devices used in education. Given the amount of educational processes which have shifted from teachers to proprietary digital technology, the last two decades have seen perhaps the greatest privatisation of education in history.

This contrasts with traditional educational data, where education was treated as a public good, and distribution of such data balanced student privacy needs against transparency as a requisite for research, and similar public goals. In Massachusetts, for example, if one wants data about a school system, one can file a FOIA request, and so long as it does not violate student privacy or integrity of assessment, the school is required to provide that data.

Current laws are increasingly proving to be ill-equipped for either supporting the potential of educational technology to improve learning, or in managing the complex privacy issues which arise with increased use of technology and new types of educational data. In the United States, the core regulatory framework surrounding student privacy, the Federal Education Rights Privacy Act, dates back to 1974.

The network of international regulations is equally dated, baroque, and often virtually impossible to follow in the era of cloud computing. The issue is not one of either too much regulation or too little – it is simply obsolete, and neither effective at encouraging innovation and progress, nor in promoting student privacy, nor in maintaining transparency of educational data, nor in maintaining education as a public good.

There are research gaps as well. Education is concerned with long-term goals, such as employability, critical thinking, and a healthy civic life. While we have plenty of data about such in educational datasets, we are still developing tools to translate raw data into meaningful insights and measurements. In most areas, the educational field primarily relies on theoretical and substantive arguments (Dede et al., 2016).

Similar reports in the United States have recognized several additional themes (Dede, 2015): (1) Mobilise communities around opportunities based on new forms of evidence (2); Infuse evidence-based decision-making throughout a system; (3) Develop new forms of educational assessment; (4) Re-conceptualize data generation, collection, storage, and representation processes; (5) Develop new types of analytic methods; (6) Build human capacity to do data science and to use its products; (7) Develop advances in privacy, security, and ethics. Of these, the deepest issue potentially is that the education sector lacks human capacity, tools, and computational infrastructure required for effective data collection, cleansing, analysis, and distribution, or even understanding the results of such analyses. Developing such capacity is prerequisite to addressing any of the remaining issues.

Key issues and Challenges

Learning at scale

Learning-at-scale grew out of several observations:

- Digital technology progressed to where it is sufficient to capture substantially all of the educational process. We can create on-line courses which are as effective as in-person ones;
- Economies of scale were recognised as necessary to enable broad-based application of evidence-based techniques in teaching-and-learning. Such techniques can lead to substantial learning gains;
- There are significant gaps in access to education. Over half of the world does not have adequate education. We can now address this problem in a way which is self-sustaining and profitable.

Learning at scale was popularised by the Stanford AI Course in 2011, from Know Labs/Udacity, and was used both for on-campus education at Stanford, and also taken by tens of thousands of students on-line. In its wake, courses rapidly followed from Coursera and edX, also jointly used in pure online and in on-campus settings. Since these early courses, learning at scale has grown rapidly. Together, these platforms represent thousands of courses from hundreds of institutions taken by tens of millions of students worldwide.

As a result, we are beginning to create data with previously unheard of breadth. Today, on the edX platform, over 500 students have finished at least 32 courses – roughly the equivalent coursework of a degree program. For those students, there are longitudinal, fine-grained data, across a broad set of disciplines. In theory, this gives the potential to study the evolution of creativity, problem-solving processes, writing processes, and soft skills over years, and see the effects of small different types of educational experience on 21st century skills. In practice, such potential has not translated into reality, and there were several key challenges.

Although large-scale educational data exists, no one has access to it. Students take courses across many providers. In the MOOC ecosystem, the market is divided between Coursera, edX, FutureLearn, Udacity, as well as a large number of small players. While these organisations have moderately large internal datasets representing at least a significant portion of a student's learning data, most research is done at universities. Universities generally only have access to data on their own students.[108] Yet the most interesting results we have seen span across multiple courses from multiple universities.

[108] This is starting to change with programs such as edX' Research Data Exchange (RDX).

In addition, we are still developing the tools and human capital to be able to manage such data. Big data is defined in terms of (1) volume, (2) velocity, and (3) variety (Laney, 2001). Data from at-scale platforms is currently in the terabytes. Once multimedia data are included (for example, student conversations over video conferences), it will move into the petabyte or exabytes scale. Education researchers are ill-equipped to manage such data, and most traditional education research is done in spreadsheets and tools like R where data fits into memory. Computing on clusters is far outside the range of expertise of most education research institutions.

The data has incredible variety. Even seemingly small changes in pedagogy sometimes lead to large differences in analytical techniques. For example, the first edX course, 6.002x, was based entirely on complex, open-ended design and analysis problems. Students received immediate feedback on their submissions, and were able to try problems as many times as necessary to achieve the correct answer. When an education research group at MIT tried to apply item response theory, a very classical and well-understood psychometric technique, it took several adjustments to make the technique work on such data (Champaign et al., 2014).

Such issues are common even within the comparatively homogeneous datasets from MOOC platforms. We have not yet developed tools or techniques for making sense and integrating the great variety of educational data from the growing range of education tools available. Traditional educational research is based on statistics. Diverse datasets lend themselves better to machine learning techniques. Few researchers at schools of education are qualified to work with terabyte-scale data, few have the skills and competencies to make effective use of machine learning, and few have access to enough data.

Such capacity is not even being developed. Organisational issues prevent most schools of education from even hiring big data or machine learning researchers. Machine learning is not considered education research, and education is not considered computer science research, so universities are not structured to hire people who work at the intersection. This gap is not widely understood or acknowledged.

Finally, such data have great velocity. The time-to-insight speed is essential. Education, as most fields, benefits from continuous improvement. Post-hoc analysis can drive policy choices. Semester-to-semester feedback can help drive improved course design. Day-to-day feedback can help instructors identify where students are struggling, and provide feedback. Second-to-second feedback allows us to provide just-in-time feedback to students themselves, to help them identify problem areas, and remedy knowledge gaps and misconceptions. Processing data in real-time at the velocity coming out of at-scale learning systems is still an area of early research with challenges difficult even for highly skilled computer scientists.

All of these areas, especially human capital development, could benefit from focused initiatives and government support.

New modes of assessment

New modes of assessment present significant opportunities for the process of student assessment, but present complex challenges to education systems. Currently, we assess how well students, instructors, schools, school districts, and nations perform through standardised exams, such as PISA (OECD, 2016d) and TIMSS.[109] Such exams are limited in time, and do not efficiently assess complex skills, such as groupwork, engineering design, or creativity on a student-by-student basis. With increased use of digital

[109] http://timss2015.org/ https://nces.ed.gov/timss/

technology, we are now collecting data about such processes, and in abstract, we have sufficient data to make accurate measurements (Mitros et al., 2014).

Indeed, there maturing research field for drawing inferences from student click-log data, known as stealth assessment (Shute and Ventura, 2013). For example, we know that experts can memorise complex domain-specific patterns, whether a circuit design or a chessboard by chunking information, whereas novices cannot. If we wish to estimate expertise, we can look at proxies during problem-solving processes such as how often such schematics were re-reviewed during the problem-solving process.

In some environments (still primarily distance and on-line education, but increasingly in digital tools used in traditional education), we can capture almost all student discourse. We can mine student interaction data for information about soft skills, and there is a growing amount of research about how to do so effectively and constructively, in societies as diverse as the Society for Text and Discourse[110], the Society for Learning Analytics Research[111], and the International Society of the Learning Sciences[112].

While we have many of the pieces in place to capture data continuously over decades of a student's education, and then visualise how such skills develop, we have not yet done so. We have not even brought together such data into one place. Once we bring together such data, we need to find ways to analyse many diverse types of data. From there, we still must to find ways to normalise across different activities which occur in different classrooms. For example, if one student performs a dozen design projects in mechanical engineering, and another in electrical engineering, we have sufficient data for both on creativity, but such data is different.

There is optimism that such problems can be solved. There are classical, well-understood solutions to similar problems in test item design (Wright and Stone, 1979). However, this will take time, and most nations lack basic infrastructure to even begin such work. For such work to begin, school systems would need to know what software is in use, have repositories for data from such software, have means (both technological and legal) to collect such data into integrate repositories, and have ways for researchers to access such data. In addition, nations would need sufficient human capital and funding to make effective use of such data.

A second policy problem is how we use results of such assessments. While more measurements and more accurate measurements have, on the whole, improved educational systems, this has come with significant costs. There are skills we cannot effectively measure. Policymakers and administrators have a strong tendency to tie incentives to student, teacher, and school performance, usually ignoring research consensus an appropriate and inappropriate use of high-stakes tests exams (Regalado, 2012, AERA, 2014). Unfortunately, as stakes go up, the accuracy of our measurements goes down. Once there are incentives in the loop – beyond simply using such data to inform teacher actions – there are incentives to game such systems.

Inappropriate uses of assessment may have significant additional unintended consequences. Students, especially more affluent ones, may take test preparation courses whose primary goal is to train students in test taking to bias their results, increasing socioeconomic advantages. Students and schools have incentives to teach to the test. Tests measurements often have errors which correlate with race, gender, or nationality (OCR, 2000). Addressing uses and misuses of such data is essential both to make good use of advances in big data in education, and to building stakeholder support from teachers, students, and voters for putting big data technologies in place. Any such

policy proposal ought to have ways to include theory around higher-level, hard-to-measure skills, and activities designed to develop hard-to-measure skills – especially those whose development would adversely affect metrics.

Data as a public good

The issue of access to data is invariably tied to the issue of student privacy. Students have an expectation that casual activities shall not be used against them in the future. In classroom design, intellectual risk-taking, the ability to make, correct, and learn from mistakes, and the ability to play are all critical. This is supported by academic literature in gamification, in psychology of motivation, in mastery learning (Bloom, 1984), and in practitioner literature on education (Kamentez, 2015). Equally importantly, in most cultures, there are beliefs about individual rights to privacy which ought to be respected.

Privacy is a challenging problem. The most common proposals, such as de-identification, simply do not work effectively. De-identification is the idea is that if we replace some identifiers, such as names and identity or social security numbers, with random numbers, and remove some additional information; such data may be safe to share. There have been numerous de-identified data sets released across multiple industries. For example, in 2006, America Online released a dataset of search data, with usernames and IP addresses removed, and Netflix released a dataset of search queries. A portion of users in both datasets were quickly re-identified by attackers (Narayanan and Shmatikov, 2007). When Massachusetts released a database of anonymised medical data, it was de-anonymised by combining with publicly available voter data (Ohm, 2009).

De-identification, while maintaining sufficient information to accurately perform a substantial portion of research on such data, is technically, and provably impossible. The most de-identification schemes can accomplish is to prevent some types of casual mistakes.[113] While techniques like k-anonymity and l-diversity can create provably anonymous datasets, once enough data is stripped out, most research becomes impossible, and where it is still possible, many research results are inaccurate (Daries et al., 2014). Worse still, educational datasets increasingly contain audio recordings, group projects, student discussions, and other forms of data which are fundamentally impossible to de-identify.

Consequently, there is a need for further building out models for maintaining physically and digitally secure access to data, and legal frameworks for deterring misuse of such data. This is mostly a policy question, not a research question. There are many well-developed projects which personally-identifiable educational data, such as Databrary[114], ASSISTments[115], and PSLC Datashop[116]. There are similar models around health information. There are newer cloud-based models where researchers may, for example, develop analyses on test data, and have such analyses run remotely on real data. It is sufficient for policy makers to pick a model and put together legal and funding support for implementing that model. A particularly good model is the Federal Statistical Research Data Center (FSRDC[117]), which is detailed in the next section.

[113] However, de-identification for prevention of casual mistakes is still important. For example, in the early days of edX, a researcher included real usernames on key nodes in a sociogram in a draft publication. Another researcher accidentally ran into data about several people they knew in real life. While a deidentified dataset will not stop a determined attacker, it can prevent this sort of unintended error.
[114] https://nyu.databrary.org/
[115] https://www.assistments.org/
[116] https://pslcdatashop.web.cmu.edu/
[117] http://www.census.gov/fsrdc

Education in 2030

One of the goals of this report is to describe how the school systems may evolve with time, so that EU policy may support (or at least not impair) positive evolution. Descriptions of possible futures are always largely speculation, and while the future is not fixed by historical forces, it is determined by choices we make today. However, speculating about possible scenarios is helpful to craft forward-looking policies, in business as in government. If policies drafted today are supportive of the many possible good future scenarios, and prevent the possible bad ones, they are more likely to work in the future.

Educational resource production, data collection, and educational technology is likely to become more unified and centralised. Historically, classrooms were fairly independent, with each teacher substantially defining his or her own curriculum. Due to progress in both education and technology, this model is already starting to change.

Blended learning has tremendous learning gains. The gap between blended and traditional is only continuing to grow, as we develop better educational models, as well as technologies to support them. Technology can manage increasing parts of the education process. A decade or two ago, online video essentially did not work, online social was immature, interactive simulations required installing software on each computer[118], students and teachers had low levels of comfort with technology, and user interfaces generally had serious issues[119].

It was impossible to centralise education in any meaningful sense, beyond basic resources like textbooks, learning objectives, and some assessments. In addition, we were not yet quite sure about what worked and what did not. While some technologies showed substantial learning gains, overall, there was no consensus about which ones, and merely adding technology to a classroom seemed to result in no appreciable gains. Consequently, teachers did most of the work individually in classrooms.

Today, much more of the educational experience is enabled or enhanced by digital technology. Active learning activities increasingly replace lectures leading to superior learning outcomes and student satisfaction. Online assessments provide immediate feedback, mastery, differentiation, and adaptivity. This too leads to superior outcomes in both student learning and engagement. While there are educational technology platforms which do not yet work well, they are fighting superior competitors, and will likely fade with time. The role of the teacher is shifting from the primary source of information to working with students 1:1 utilising such digital materials.

The question, then is, where the blended resources come from, and where data about their usage goes. Economics drives curriculum, course, and educational resource design to be centralised. It's a natural monopoly, and there are fixed costs to creation, and near zero incremental costs to additional usage. It's not just a natural monopoly – it has strong network effects. A platform with more students and teachers has access to more data, to more contributions from teachers and students, and to a more diverse group of students. Student forums have more activity.[120]

[118] Technologies like HTML5 and Javascript were not yet mature enough to handle high-quality educational experiences.

[119] In part because the buyer wasn't the teacher or student, but a school administrator. Consequently, purchasing decisions focused on issues such as authentication, administration, rather than student experience.

[120] In 6.002x, we saw students respond to 92% of questions, with a median response time of 12 minutes, and answers far better than a traditional TA, in part due to having over 7000 students who were active enough to earn a certificate. See MITROS, P. F., AFRIDI, K. K., SUSSMAN, G. J., et al. 2013. *Teaching Electronic Circuits Online: Lessons from MITx's 6.002x on edX*. Proceedings of the IEEE International Symposium on Circuits and Systems (ISCAS), Beijing, China. Published May. Available: Mitros 2013/IEEE 6.002x paper. [Accessed March 2 2017]..

All of this leads to better learning outcomes. A platform with just a few thousand students can crowdsource interventions for common student errors, to outperform intelligent tutoring systems with multimillion dollar development costs, for free (Mitros and Sun, 2014). For connecting students to jobs, students will want to go where employers are, and visa-versa, and a newcomer has a chicken-and-egg problem.

Businesses are starting to understand that education is big business and to invest in this space. Governments spent about $3 trillion on education annually.[121] The education gap – how much wealth would be created in the world if we educated everyone – is in the tens of trillions of dollars annually.[122] That's a huge business opportunity. Major corporation such as Facebook[123] and Apple[124] have major education initiatives. Investors have financed the three major MOOC initiatives, Udacity, Coursera, and edX, at a level of over a third of a billion dollars, and their valuation is many times that.[125]

These are not things which would have happened in the educational technology space a decade ago. The technology and the learning gains were not there, and investment in education was either viewed as philanthropy, or as branding and customer development.[126] The pace is very fast, compared to traditional education or policy. edX is a half-decade old, and has over 10 million learners, 100 partners, and 1000 courses. Major announcements come from the MOOC players monthly, whether a new partnership with a government, or a new accredited online program.

In many ways, the landscape of educational technology resembles that of computing circa 1975, or e-commerce circa 1999. There are many competing platforms, and while it is too early to tell which ones will dominate, it is clear that in the coming decade, we will see the Amazon's, eBay's, and Microsoft's emerge. How this centralisation happens is important. Microsoft did not innovate in its core business from 1995 through 2010, until it was threatened by cell phones and the Internet[127]. eBay hasn't innovated substantially since around 2005. Amazon is aggressive and innovative moving into new markets, but the basic online store has functioned with little progress over the past decade.

Without appropriate regulation, if a monopoly is in place, progress generally stops. In contrast, Wikipedia, a monopoly careful stewarded for the public good, has continued to grow and improve continuously since it's founding. Apple and Google have a duopoly on cell-phones, which has led to rapid innovation. Social networking has formed a set of relatively narrow segments, each dominated by a different player, such as Facebook or LinkedIn. Since each controls a well-defined segment, while there are multiple players, they essentially do not compete. Neither platform has improved had a particularly fast rate of improvement.

We are at the point where policy can help decide which of these models comes to dominate.

---

[121] http://data.worldbank.org/indicator/SE.XPD.TOTL.GD.ZS

[122] 3.7 billion people earn under $1500 per year, according to https://www.givingwhatwecan.org/get-involved/how-rich-am-i/?country=USA&income=1500&adults=1&children=0. Education increases productivity to where, outside of active war zones and similar areas, an individual with education on par with a US Bachelor's Degree rarely makes under $10,000 per year. This analysis omits improved productivity due to reduced corruption and other ancillary effects which increase productivity. It also omits resource limitations, and increased competition, which would reduce the gains.

[123] http://www.theverge.com/2015/9/3/9252845/facebook-education-software-plp-summit

[124] https://itunes.apple.com/us/app/itunes-u/id490217893?mt=8

[125] edX is a not-for-profit, initially funded by university endowments.

[126] In particular, several graphic design and engineering organization took control of markets by providing subsidizing education initiatives, such that professionals entering the market were familiar with their tools.

[127] It is pretty astonishing that large computers were outclassed by tiny devices, with a fraction of the screen size, processing power, or storage. Much of this is the result of the failure of Microsoft's technology to make effective use of high-resolution displays, personal servers, and many other technologies commonly used in Unix systems.

Implications for EU policy

Regulation at the country and the education system level is mismatched to the growing use of educational data, especially to the growing range of uses which are increasingly social and data-driven. As we see a shift towards continuous formative assessment, group projects, soft skills, cross-cultural skills, adaptive assessment, data-driven policy, stealth assessment, and discourse analytics, existing policies often become ambiguous and impractical.

As educational technology moves to support 21st century skills, there are increasing amounts of data which are common to multiple students. For example:

- Where educational work is jointly created by multiple students, and the system maintains a log of all contributions, it is impossible to remove the contributions of one student without affecting the other students;
- If there is a threaded forum discussion, with student back-and-forth, it is, again, impossible to forget one student while maintaining the contributions of the remaining students who might respond to comments by one student;
- Adaptive systems build models based on students' actions. Forgetting the actions of one student may break such systems.

Given such data, individual rights, such as the right-to-be-forgotten, are increasingly impractical. A right is only helpful to have if it is also practical to exercise that right. However, if a significant number of students request to be forgotten, that right would either be incomplete (if such shared data were not removed), or would break the experience of other students (if it were).

In addition, data-driven decision making would suffer. We can expect the set of students who opt out to have significant bias: for example, students who made embarrassing mistakes would be more likely to want to be forgotten. This would adversely affect the replicability of studies, introduce biases into analyses, and so adversely affect any decisions made based on such data.

Instead, policies should focus on how such data is used. Data securely stored, and used for no other reason than to the students' benefit poses little risk. At present, few jurisdictions have legal requirements that such data be stored securely, and there is little transparency to how it is used. Terabytes of fine-grained student interactions are being captured by proprietary vendors.

Security practices around such data are based on limiting commercial harm, and use of such data is based on commercial benefit. There is a general consensus among educational research that traces from such educational data ought to be treated as a public good to the extent possible while remaining faithful to student privacy, but this is increasingly not the case.

However, in virtually all cases, such data are not under the control of the students about whom such data is collected. It is not portable between systems. If a student begins their education with one vendor, they are locked into that vendor as that vendor begins to builds a model of that student's knowledge model, social network, and a collection of that student's work. The student has no way to inspect such data. It is also unavailable to the student's teachers, or for scholarly and policy research.

Educational data are of limited use if they are distributed among many independent data stores. To realise the potential of such data, data must be correlated across courses, longitudinally, across students' entire lives. We are starting to see the power of pivotal studies which look at effects of education decades later (Sass et al., 2016, Dobbie and Fryer Jr, 2016). Thus far, the measures these studies use are coarse. Those examples analyse the effect on income of charter a quarter of a century later.

With modern educational data, we can compare educational approaches used in individual classrooms (for example, intelligent tutoring systems versus project-based learning on long-term outcomes), outcomes for individual teachers, and where there is statistical significance, individual course resources and assessments. In many ways, this is the silver bullet of educational policy research – right now, we are largely limited to very simplistic proxies for many types of skills.

There are models which provide full transparency for research and policy reasons, while fully preserving student privacy. Perhaps the best established is the FSRDC model. In this model, the government runs a set of data centres where individuals and researchers may access full, uncensored data (with obviously identifying information obfuscated to prevent casual errors). Researchers may not remove anything from such a research centre, except aggregate results. Under this model, all ed-tech providers serving EU classrooms would be required to deposit their data in such a data centre. These centres would be available to the general public for supervised access. In addition, students would be able to export their own traces, as well as make them available to their instructors, as well as other educational technology providers.

To make sense of educational data, context is critical. Student data traces make no sense without understanding the context in which they were generated. A regulatory standard ought to promote both open standards (e.g. LTI Caliper and xAPI for such traces), and the use of free/open source educational software, or at the very least, software which may be inspected and used by educational researchers at such data centres.

For student privacy, there should be thoughtful consideration to how such data ought not be used. This question has no uniform solution. It is dependent on the culture of a nation, and the beliefs about privacy of that nation, as well as on the economics of each country. Such data are increasingly lucrative for job placement, and helpful for college placement, national security, and law enforcement. To what extent such uses should be permitted ought to take into account the beliefs of people in a country.

Policy uses may have adverse effects as well. We can extrapolate tremendous amounts about the abilities of individual students, teachers, schools, and school systems based on data already being generated. However, most such inferences are based on correlations, not causality. Once inferences drawn from such are used for high-stakes, whether for individual student placement, teacher pay, school pay, or otherwise, they are stop being good measures; they are liable to be gamed.

This has been seen in high-stakes testing in the United States, where exams focus on simple measures, such as memorised vocabulary and basic algebra skills, and consequently, classrooms focused on those to the detriment of higher-level soft and communication skills, quantitative reasoning, and mathematical maturity. Measurement, simply presented as an indicator to teachers and students, rarely causes harm, but great care must be taken when closing the loop to assessment with any sorts of consequences.

As educational data become increasingly international, navigating the international legal landscape becomes complex. An educational technology organisation operating globally must comply with legal regimes in approximately 200 countries in the world. To give one example, when MOOCs launched, the State of Minnesota responded by banning providers from operating in the state (Mangan, 2012). The US Department of State enforced trade embargoes against students in Cuba, Syria, Iran, and Sudan (Straumsheim, 2014).

The US Department of Justice brought forward an Americans with Disabilities Act enforcement action, to require WCAG compliance, captions on all videos, and high colour contrast in graphic designs (Cooper, 2016). Such enforcement actions, while generally individually reasonable, together cost millions of dollars to respond to, and effectively

limit the MOOC ecosystem to players who have budgets of tens of millions of dollars. This dramatically reduces competition and innovation in this space.

We do not propose standardising such policies by international treaty. A diversity of regulatory approaches is healthy, especially in young, dynamic industries where it is not yet clear which approaches might work best. However, much of the current diversity is accidental, unintentional, and unnecessary. Often, two countries will have policies which are substantively identical, but differ in detail sufficiently to require twice the legal cost. The licensing of creative works offers a helpful model of how unnecessary diversity may be constrained. The vast majority of openly licensed written works are available under a Creative Commons license. This framework provides a basic set of pluggable license terms which authors may compose to reasonably approximate the set of licensing restrictions they would like.

Likewise, most free and open source software uses one of a diverse (but constrained) set of licenses. Both common free software licenses and CC licenses are well-understood. Some authors add additional restrictions, which are easy to find as well, as changes from a standard. A regulatory regime which had draft language of terms, and clearly documented variations there-of would be helpful to allow smaller providers to understand the requirements in each jurisdiction, and to compete without astronomical budgets. In other words, even if there is a right-to-be-forgotten, unless there is compelling reason to do otherwise, there should be just one such right, with common mechanisms for notification, response times, and otherwise, in all jurisdictions which choose to implement it.

Student tracking

Introduction and context[128]

This section surveys the opportunities and challenges afforded by the use of student tracking with big data to inform educational policy and practice in Europe. The Europe 2020[129] strategy aims to stimulate a competitive and economically viable Europe. Central to this vision is a society where European citizens are given the best educational support possible for their individual learning needs. To achieve this vision, there is an ongoing discussion as to whether the education and training sectors across Europe have to shift their objectives and structures. Educational tracking, utilising largescale datasets known as 'big data', has been proposed as one solution that would allow for more tailored educational pathways.

The term student tracking has been used inconsistently in official documentation and reports. For example, Gaebel (Gaebel et al., 2012) used the term to refer to a 'system of tracking student progress throughout their educational lifecycle'. However, in writing this section we were encouraged to take a narrower taxonomic standpoint. Thus, we will follow the definition of Hallinan (Hallinan, 1994) which refers to the process of classifying students by ability, and then separating them, for example teaching them in 'ability groups' or to be taught separately (such as in different schools, buildings, rooms, or social circles).

The relationship between forms of tracking used to inform educational policy and practice is important, and micro-level behavioural observations can be used to transform practice in ways that achieve macro-level policy objectives. Ultimately, understanding student behaviour and problems with learning can feed back into policy decisions around where and when students should learn specific skills. Thus, these innovations have the potential not only to improve the efficiency, speed and accuracy of policy forecasts, but also to transform the educational practices that underpin policy implementation. The key to the success of this vision is the measurement and use of appropriate big data.

Education policy work can be guided by datasets that link different types of data. For example, UK Government data is now combined with labour market data to pinpoint which types of expertise are in demand by employers. The outcomes are fed back into educational policy to ensure that the education system teaches the skills needed by companies (DfE, 2016). This innovation supports education sectors in responding to economic needs in a more agile way. An important question is: should we track groups of students with different abilities and stream them into different educational trajectories based on predictions of future labour market's needs?

Until recently different sorts of data were gathered and used for different purposes and held in separate databases so were difficult to combine. To alleviate this problem EU governments and intergovernmental agencies, such as the Organisation for Economic and Co-operative Development (OECD), have been prioritising making macro-level datasets openly available to support policy development informed by big data: for example (Britton et al., 2016, Britton et al., 2015, Holland et al., 2013).

Nonetheless, student tracking is a controversial practice within education. On the surface, separating students into groups based on academic ability appears to be an idea that leads to many benefits; in theory teachers can target the difficulty of learning materials and tasks to all members of class, rather than setting work that will be too easy for some and overly difficult for others. A 2007 report by McKinsey (McKinsey,

[128] Thanks to Victoria Murphy and Vasudha Chaudhari, Open University, who contributed to the ideation and writing of this section.

[129] https://ec.europa.eu/info/strategy/european-semester/framework/europe-2020-strategy_en

2017) examined which countries were succeeding in the Programme for International Student Assessment (PISA) to understand what common factors the top scorers possessed. One of the aspects of good educational systems highlighted was that each child should receive quality instruction. Tracking is one potential way that schools can support students in receiving guidance that is aimed at the right level for them. Tracking has been suggested to help high performing students to challenge themselves, being stimulated by being surrounded by those of equal ability, and to discuss and explore learning opportunities (Fiedler et al., 2002, Kulik and Kulik, 1992).

The potential of tracking systems can be seen from those countries that currently embrace them. A good example is that of Singapore, who were rated number one for Maths, Science and Reading in the 2015 PISA results. It is ranked first for Maths and Science of both 4th and 8th grades according to the 2015 Trends in International Mathematics and Science Study assessments, and is a top performer according to the National Centre on Education and the Economy's Center on International Education Benchmarking.

At the end of primary education students take an assessment which will decide the type of secondary school that they enter, and ultimately the kind of qualifications that are received. Multiple schools in Europe have also applied some variety of tracking. The Netherlands, for example, has an education system that utilises streaming from a young age and is ranked among the best of European educational systems according to international assessments (PISA). Similar to Singapore, an aptitude test is administered at the end of primary education to guide teachers and parents in recommending what type of secondary education to pursue.

Tracking has, nevertheless, many opponents who disagree with streaming of students by abilities. The aforementioned report from McKinsey and Company (McKinsey, 2017), for example, outlines that successful education systems set high expectations for all students, rather than just those who are academically gifted from a young age. The 2012 OECD Equity and Quality in Education report found that ability tracking often widened the achievement gap of the highest and lowest students (OECD, 2012). This is supported by multiple empirical studies. Hattie additionally found that while ability grouping did have a small positive effect on student achievement, it was one of the least effective approaches to increasing student capabilities (Hattie, 2015).

Hanushek and Ludger found similar results in their study comparing on an international scale systems that use tracking and those that do not (Hanushek and Wößmann, 2006). While their study concluded that the inequality gap in terms of student achievement was consistently worsened in educational systems using tracking, they failed to find evidence that this was associated with an increase in average achievement of students. At its worst the implementation of a tracking system runs the danger of negatively impacting the majority of the student body, due to polarisation of top and bottom students, high achieving students being forced to advance at a rate that is too quick, and a lack of pedagogical variety due to perceived homogenous classrooms (Boaler et al., 2000).

Ultimately student tracking is one of many potential ways to improve student performance. However, careful consideration must be employed by policy makers in the creation of a system that utilises tracking as the evidence for its effectiveness remains minimal. The use of big data to inform algorithms could more suitable suggest paths for students, but the effectiveness at creating educational equality and actually improving results should be carefully monitored.

The tracking systems that currently exist in the EU and seem to be the most effective allow for students to choose to follow the path that suits them the most. For example, in Finland students can choose to follow an academic or vocational track for upper secondary school depending upon their interests. Big data that supports better tracking could generate recommendations that take into consideration far more factors than

current systems, but ensuring that students are active agents in their educational choices is an important aspect of encouraging individuals to take a life-long responsibility for their own choices. Key in this, however, is that better tracking also supports better student learning mobility, and that students are not 'de-selected' early in their education and then remain as low attainers

Big Data

Big data has been defined as *"the Information assets characterized by such a High Volume, Velocity and Variety to require specific Technology and Analytical Methods for its transformation into Value."* (De Mauro et al., 2015, p.103). The impact of big data on education is governed by characteristics that contribute to improved analytical ability - volume, velocity, and variety.

Volume in this context means the massive amount of data available online. For example, it includes performance statistics, student records, data from online learning platforms, and how long a student spends on a task. The volume of data created is increasing rapidly, due to the escalation in the number of learners accessing Open Educational Resources (OER) and Massive Open Online Courses (MOOCs). According to the ICEF Monitor, enrolment in MOOC platforms alone surpassed 35 million students in 2015 (ICEF, 2016).

With this increase in the number of learners comes an expansion in the rate of generation of data. *Velocity* refers to the rate of data creation. One example of high speed data creation is 'clickstream data' which are generated as students interact with course platform tools, content and with their peers. Data generated in the form of text messages (including Twitter feeds), images, or audio content during these interactions may be collected, processed, and stored in a meaningful way to allow for intelligent analysis. Data is generated rapidly and requires specialised technology for storage and retrieval.

Variety refers to the diversity of big data sources, such as images, textual data, tweets, and click-stream data. Data from these sources are often unstructured, and even within the same source varies widely. For example, two emails could differ in terms of their length, attachments, colour of fonts, inclusion of recipients and in many more ways. Techniques used in big data analysis must be able to deal with this variety of forms.

Despite this promise of the use of big data to support policy and practice work, success has been more difficult than anticipated, and there are a number of problems limiting its use. These problems are related to the nature of the data and, as discussed later, influence the effects of monitoring of educational systems and student tracking using big data.

Some problems are associated with the sourcing, storing and analysis of big data. This is particularly difficult where data sources are distributed across multiple sources and servers, and are gathered using different methods. To enable comparability, data must have standardised quantitative and qualitative indicators that offer insights policy and practice.

Others issues relate to the assumptions underlying the interpretation of the data, sometimes over-simplifying complex processes. Gathering micro-level student data is complicated, not only because the variables are complex, but also because it requires intensive data-gathering and real-time analysis. The available data is limited to online learner activity and do not offer direct evidence of offline activity and cognitive development. Also, there are important ethical implications associated with the use of student data (Slade and Prinsloo, 2013), such as transparency, consent, and rights to seek redress.

Therefore a major challenge for Europe is to implement student tracking in ways that

enable us to extract meaning from large datasets being generated through micro-level, online student activity and to distil this data into usable and equitable tracking information for students, teachers, and governments (Dede et al., 2016).

Societal trends influencing student tracking with big data

Developments in the use of student tracking with big data are framed by a range of broader societal trends:

Abundant social data and algorithms. The Europe 2020 strategy acknowledges the increasing ability of machines to deliver impartial, intelligent decisions through algorithms that analyse large amounts of data. An algorithm is *"any well-defined computational procedure that takes some value, or set of values, as an input and produces some other value, or set of values, as output. An algorithm is thus a sequence of computational steps that transform the input into the output"* (Cormen et al., 2009, p.5). With the ability to capture data through the 'Internet of Things' (the interconnection via the internet of computing devices embedded in everyday objects, enabling them to send and receive data) all EU citizens are subject to algorithms in almost every aspect of our daily lives.

An acceleration in the use of 'smart' technologies in everyday life, such as mobile and wearable devices that gather personal data offer opportunities in terms of where data can be sourced to support student tracking. A trend towards increased sharing of data has become routine through reviews and recommender systems for services. These sorts of activities impact societal expectations of the systems that support education. It is almost routine for students to voice opinions about their education through social media, for example Facebook, Instagram, Twitter, and other digital tools. An important question is 'What would a tracking system that incorporate pupils' and students' social media data look like?'

Data ownership. As data become more abundant, and boundaries across more organisations become permeable, large datasets can be under the control of a range of different stakeholders. For example, a digital profile is now routinely used to access multiple social platforms, such as the sign in information for Facebook or Google accounts. There are also instances of students being asked to provide data to an institution for specific purposes, such as teacher assessment systems: for example, see (Rahman, 2013).

These applications can be controversial, since the criteria used for assessment by students might not align with the institution's assessment criteria, but, perhaps even more importantly it is the students - or the system provides the student uses - rather than the institution and state, that control the data. This issue raises interesting discussion on legislative frameworks. It is important, therefore, to consider 'who owns the big data that is the product of student monitoring and educational tracking?' And what effect might different ownership models have on future monitoring of educational systems?

Permeable education boundaries. The idea of an education and 'job for life' has reduced significantly. People now expect to weave in and out of careers and education pathways, a phenomenon termed by Arnett as 'emerging adulthood', particularly for young adults who wish to explore and try out different career pathways (Arnett, 2000). A key point here is that 'learning' is not limited to formal education settings, but also spans informal contexts. The use of data therefore needs to be accompanied by a rethink of the potential of data in education, since limiting data gathering to formal education and courses limits the impact of big data and tracking. This situation presents challenges as to how students can be tracked. An important question to consider is 'what kind of educational opportunities does big data tracking support?'

Ethical considerations. The ethical implications of using data for student tracking are

complex, bearing in mind the relationship between accessibility and availability (boyd and Crawford, 2011), and privacy issues are considered in another section in this report. Current work on student monitoring, has highlighted the need to define the context and extent of tracking[130]. Recent debates have focused on whether students should be able to opt out of tracking, particularly if they benefit from the tracking of other students' data (Slade and Prinsloo, 2015).

There are unanswered questions around whether it is legal or ethical for schools to widen the scope of data used to track students, for example use their social media data for analytical purposes. Thus it is critical to consider 'how should ethics be framed in terms of big data in education, with a focus on monitoring'?' There are issues over privacy and transparency at the school level that have yet to be resolved. For example an important question is should teachers be able to view all student data? Seeing data allows teachers to know when some students need more interventions, but there is also a need for students to be able to control privacy settings and to 'opt in and out' of data tracking as appropriate.

These trends frame key challenges around the use of big data for student tracking, which are explored in the next section.

Key issues and challenges in the use of big data for student tracking

What kinds of educational opportunities does tracking support?

One major criticism of tracking is that if it is applied too early those from poorer or immigrant backgrounds will be particularly disadvantaged (OECD, 2012). Studies have shown that even from ages as young as two, those from families of lower socioeconomic status test significantly lower on supposedly standardised tests, such as IQ tests (von Stumm and Plomin, 2015).

Although big data could be used as a method to ensure students from disadvantaged background with the potential can strive if streaming occurs early in a student's timeline, there is more evidence to support streaming occurring later in life (Lavrijsen and Nicaise, 2015, Pekkarinen et al., 2006). Delaying tracking decisions also has the advantage that more data on each student can be collected to better inform any algorithms used in tracking based on big data.

Another concern that may arise from ability streaming is that it ensures students are surrounded by those most similarly minded to themselves. Some of the most modern pedagogical approaches are based on ideas that understanding of topics comes from not absorbing information from a teacher, but rather through discussion with those who see things from a different perspective. These pedagogical approaches have been supported with substantial empirical evidence (Mercer et al., 2004, Mercer et al., 1999).

Student tracking has the danger of creating groups from very similar backgrounds. Using big data to inform tracking decisions could potentially reduce the homogeneity of classrooms, especially if data besides academic tests is used to create groups, utilising information about hobbies and languages spoken for example. The use of technology to inform decisions does not necessarily lead to enhanced development, and there is a need for guided development of tracking and monitoring systems that will support learning in a pedagogically-based way. This micro-level use of big data to support learning is a promising use of tracking with big data.

Big data tracking systems hold much potential, additionally with regards to comparison of international educational systems. PISA is arguably the most consulted measure of the merits of different countries' educational systems, and has received criticism for overly-

[130] https://digitalstudent.jiscinvolve.org/wp/data-service/

simplifying the complexities of complicated and culturally situated environments (Meyer and Benavot, 2013). A tracking system based on big data could yield a more nuanced data set for international policy makers to better understand the effects of policy changes in terms of the demographics of individual students who end up in certain tracks, or even just in terms of educational achievements and gains.

This data could become an important source of information on which to make and evaluate EU level policies. Caution must be used, however, when assessing success purely based on academic achievement. An EU funded report from EPALE (Electronic Platform for Adult Learning in Europe[131]) identified a range of contexts where use of big data to track students across EU countries has not led to improved outcomes, largely because the normative measures of success in formal education may be very different when applied to open education and learning.[132]

Can we rely on algorithms to be unbiased when classifying students into different academic tracks?

Algorithms are developed by coders to analyse data in a meaningful way. These can be helpful in understanding data, but inevitably are shaped by underpinning assumptions and biases. Data gathered and analysed by algorithms are limited by the expertise and assumptions held by those people who write the code (Williamson, 2015a). If the coders do not appreciate the underlying assumptions of their codes, then the data the algorithms analyse can be compromised. According to Boyd and Crawford:

> "*As computational scientists have started engaging in acts of social science, there is a tendency to claim their work as the business of facts and not interpretation. A model may be mathematically sound, an experiment may seem valid, but as soon as a researcher seeks to understand what it means, the process of interpretation has begun. This is not to say that all interpretations are created equal, but rather that not all numbers are neutral*" (boyd and Crawford, 2012).

Thus, if algorithms are used to track student progress, then algorithm bias will distort how students are viewed through tracking systems.

Algorithm bias is exemplified through school games that are used to enact government policies. Games act as part of governmental processes that develop policy through the surveillance of 'psychological characteristics' of students (Williamson and Facer, 2004). These games can have interventions intended to modify attitudes, beliefs and personality through the imposition of positive affect. Thus, the social activities and rules of each game set important messages for children and define the sorts of data that are gathered within the gaming environment.

This combination of games and tracking through the use of big data amounts to an acceleration of governmental behaviour change programmes in schools. Another illustration of algorithm bias is the measurement of participation and completion in Massive Open Online Courses. The algorithms developed to track students are based on assumptions that active participation and completion are necessary condition for learning in these courses. However, these indicators do not measure learning (Littlejohn and Milligan, 2015).

Biases implicitly or explicitly are encoded into algorithms, moving away from the idea of big data models being objective and neutral, as previously highlighted. A key question to consider is 'How might algorithm bias affect tracking systems and what could be the social consequences?'. There are many ways in which the objectivity of algorithms can be comprised:

[131] https://ec.europa.eu/epale/
[132] https://ec.europa.eu/epale/node/29206

Inclusion and Exclusion. Algorithms "learn" their intelligence from the data available, so implicit biases stem from discrepancies in underlying data due to missing or biased data. Data is only available from individuals who are active online. So, as algorithms develop intelligence from the data available, decisions made by these systems will tend to favour those who are already online (since their data is being used to 'train' the algorithms). Eurostat internet statistics for 2015[133] report that 17% of European citizens are totally offline. This has long-term impact on ability groupings made through big data tracking techniques, as the algorithms that are biased towards the 'data-rich' demographics will not be as efficient in making decisions for the academic tracks or career paths of the relatively data-poor population. This situation leads to a deeper digital divide, with those already online being favoured and those who are not online being excluded.

If algorithms are used to open access to learning opportunities, improved job prospects, and effective social interactions, digital exclusion means large proportions of society remain underrepresented with respect to their needs and expectations. Cathy O'Neil warns about the perils of digital exclusion, asserting that algorithms used to analyse big data "*tend to punish the poor*" (O'Neill, 2016, p.8). To ensure inclusivity it is critical that the policies on student monitoring take into account data from people from all groups in society.

Recent advances in machine learning have seen the emergence of Embedding Technique, where each word in a vocabulary is assigned a vector, and word associations are formed between related words. This method allows social media sites such as LinkedIn to make recommendations. However, the traditional gender biases tend to view women specific roles – such as nurses, receptionists, or teachers, rather than CEOs, investment bankers or consultants. This bias is exacerbated by the embedding technique. Similarly, other algorithms reflect common societal biases.

Data openness (or transparency) does not negate biases, due to complexity and opaqueness of learning mechanisms. One attempt to overcome this problem is led by Google research scientist Moritz Hardt. His team has been working on a vetted methodology to reduce biases related to gender or race introduced into learning algorithms (Hardt et al., 2017).

Educational policies that reduce algorithmic biases must go hand-in-hand with ethical discussions around the fairness of these algorithms. It is important that the trends discussed in this section should be addressed by EU policy actions.

Correlation not Causation. A significant challenge associated with big data algorithms is that they provide elaborate patterns of correlation, however this may be often mistaken as evidence of causation. There is an energetic debate amongst data scientists arguing over the sufficiency of correlations based on huge amounts of data versus the need for finding statistical causation (Calude and Longo, 2016).

For example, big data may indicate that students who were allowed to regularly bring their own devices to school are more productive, and policymakers may look at this information and decide to implement BYOD policy. However, the correlation does not necessarily mean that bringing your own devices increases productivity. It could be that the students included in this data set are proficient in the device usage thus displaying productivity in tasks, which is lacking in students are not adept at using digital devices. Albeit being an hypothetical example, this shows that if policymakers do not keep in mind that the existence of a relationship does not prove cause, they are at risk of wasting resources to implement solutions that prove ineffective.

[133] http://ec.europa.eu/eurostat/statistics-explained/index.php/Digital_economy_and_society_statistics_-_households_and_individuals#Internet_access

Algorithms cannot capture implicit traits. Big data tracking may be efficient in assessing tangible measures such as performance, grades, attendance etc. However, educational tracking should extend beyond tacit measurements to include complex traits such as creativity, critical thinking, problem solving abilities, innovative capabilities etc. This is particularly significant if an objective of student tracking is to guide learners towards their most suited careers. Reliance on big data tracking techniques may lead to loss of vital student information.

Missing Legacy Data. Algorithms that used to analyse and convert unstructured institutional data to meaningful insights, are generally trained on the basis of near real-time data. The legacy data of the educational institutions may contain valuable information, which is lost when using such algorithms.

Can we capture the data needed for an unbiased tracking system?

As social media becomes an integrated part of EU citizens' lives it is, perhaps, natural that the social data that it generates becomes an integrated part of educational experiences. Thinking about the ways that it could be beneficial for those in secondary and tertiary education offers many ways to enhance classes and projects, from the relatively superficial such as a teacher being able to check on students' interests and hobbies to frame their lectures in a relatable way, to the transformation of communication between groups working together. Already this is being observed in schools today, but is mainly left to the initiative of individual teachers or students.[134]

Social big data, utilised by a tracking system that is integrated with social media, also has the potential to help individuals and students in ways outside of the classroom. Cyberbullying, for example, is a problem that has been recognised by the European Parliament (Dalla Pozza et al., 2016) and has been in several news headlines in recent years (Wakefield, 2017, Broomfield, 2016). Teachers and schools do not have the time to follow each of their students using social media, but big data based tracking would make it possible to have recommendations about which students are potentially the targets of cyberbullying. This could be taken into consideration in addition to academic achievement when recommending students for different tracks.

Recent research has shown cyberbullying is often the result of students intending to make a joke, or failing to make a connection between their online actions and real life consequences, leading to situations where a teacher's subtle intervention could help (Sabella et al., 2013). Big data could be used to help teachers to understand several issues that relate to students outside the classroom. Many wearable technologies offer insight into health and sleeping patterns, GPS on mobile phones could be used to know when underage students are going out of school at lunchtimes, or on visits. While tracking using big data has the ability to group students by academic ability, it also has the potential to help teachers understand what issues her group of students are facing besides academic difficulties.

While there are many advantages of using social big data in education there are an equally large number of potential moral and ethical issues. While some social data are available freely to all, such as Twitter feeds or Tumblr posts, how much should students' Facebook or SnapChat posts be available to schools? Although these tools have public sections, the privacy settings on these applications allow users limited control of their data. Often this control is insufficient or not sufficiently well understood by the user.

Therefore a key question is whether students should be allowed to 'opt in' to provide only the data that they are comfortable with giving, or will data become less useful as students become more cautious about what they post, knowing that the school or college

[134] http://www.mheducation.com/blog/thought-leadership/how-social-media-can-help-students-study.html

may be monitoring it. There are many tangible benefits that are technologically possible if schools use big data created by students' social habits, however, government or institutions who look to use this need to carefully consider the exact aim that they intend to achieve, as well as the ethical dilemmas that it may present.

In line with these concerns over the availability of social data there is also an issue with the amount of data that would be made available through a tracking system. The EU in general promotes open data[135] and open education resources[136], but the data used to inform tracking could become a valuable resource for others to take advantage of students. As the 'Right to be Forgotten'[137] has been needed, in general, on the Internet, something akin to this would be needed for the data informing a tracking system. Data transparency and control is therefore an issue that policy makers should consider.

Does the ownership of big data influence its use?

The proliferation of online data raises an important question as 'who owns big data?' (Ruppert, 2015). This question is complicated by the variety of different methods used for data collection. Governments no longer have the control of data they once enjoyed. Companies and education institutions invest in the collection, storage, and analysis of this data, with an intention to use it for mutual benefit. Questions over which data should be available for tracking decisions, who owns this data and how it should be made available are complex.

Data ownership refers to the data governance process that awards the data owner the rights to create, modify, share and restrict access to the data. Typically, a data owner also holds the right to confer these privileges to third parties. There are several advantages and disadvantages for either of the options. For instance, if the data ownership rights are awarded solely to the educational institution, then there is a risk of student's personal data being shared with third parties for monetary or business benefits.

Student tracking data could be construed as being private and personal data, therefore the default ownership of the data could be with the individual. In a ruling of the Court of Justice of the European Union, an individual's right to be forgotten was upheld over a company's property rights, even though the company had made an investment in collection of that data (Commission, 2014c). However, since student monitoring data is critical for intelligent tracking systems, the ability to 'opt-out' can compromise the support offered to students who elect to 'opt in' to having their data tracked.

Sources of big data in education include:

- Administrative data held by governmental organisations. These include longitudinal studies, cohort studies, and digital records from schools and colleges. These datasets support quantitative analysis to support policy work and learning design. The ETER (European Tertiary Education Register) dataset, UOE database, and Key Data series in the Eurydice Network are examples of administrative datasets available across the EU;
- Institutional Administrative Data. It is now routine for educational institutions to maintain digital records of their students' demographic information, socio-economic status, academic grades, and attendance data. However, the proliferation of digital technology has enabled the capture of seemingly harmless, yet highly personal data on students. For example, information regarding library check-ins, time spent on school computers, Google search histories, and timings at the cafeteria, can be used in conjunction with traditional student tracking data

[135] https://ec.europa.eu/digital-single-market/en/open-data
[136] https://www.openeducationeuropa.eu/en
[137] http://ec.europa.eu/justice/data-protection/files/factsheets/factsheet_data_protection_en.pdf

to reveal fascinating patterns about any large-scale entity (schools, universities, or any other types of educational institutions);

- Student Data from online learning platforms. Online learning environments are becoming commonplace. Data that can be recorded within these learning platforms include every keystroke, login time, quiz response time, quiz performance, and so on. This data when analysed can provide valuable insights into learning patterns of students, and can direct the creation of personalised learning plans. However, this data can be misleading since it gives a narrow view of all activities and interactions students engage in online and offline or within and outside the classroom.

Future policies governing the ownership of educational data, need to reflect dynamic developments in learning platforms, and to be fully in line with data ownership issues and data protection regulations. Even though manipulation of personal data requires advanced technical skills, student tracking and monitoring systems must have data security built in so data analysis cannot be compromised. These issues of data ownership and protection have profound ethical implications, and are explored in detail in the privacy section.

The following scenario illustrates a scenario of a typical day in a school where technology solutions are implemented for tracking:

Julia's school in Gdańsk is equipped with a state-of-the-art cloud-based school management system that integrates academic reporting, attendance tracking, curriculum management, and also includes unique features such as behaviour management, library services, cafeteria management, facilities management, and a student and teacher portal. While planning the module Julia plans draws upon external resources and advice from other teachers.

This action generates a list of external resources and all Julia's interactions are logged. On her way to class, she decides to stop for a coffee at the school cafeteria where she pays with her RFID (radio frequency identification) using her school id card. In the classroom, she logs into the school portal to access her module. The time and duration of this session is recorded.

Meanwhile as Natalia, Jakob, and Piotr work on their website project by collaborating with external technical expert Josan. Since most of these interactions are outside the digital monitoring system of the school, any library check-ins or searches are not recorded in the school's digital records. When this data is analysed it might seem as if Natalia, Jakob, and Piotr completed a project without referring to any academic resources.

This scenario reveals the extent and depth of student and teacher data that can be available within a smart school administrative system. Associated meta-data on interactions with school systems, or geo-location data may be analysed to gain meaningful insights on behavioural and learning patterns. However, other data from outside the school system is needed to offer a complete illustration of the learning behaviours. The question of who should own this data is difficult to answer. Should ownership rest with the institution that has invested in the system infrastructure and data gathering, or with the teacher or the students whose data are being collected?

How should ethics be framed in terms of big data?

Capturing and using big data for student tracking raises a number of ethical dilemmas. To utilise data that can help provide an holistic view of student progress, these data should be measured beyond the classroom. Consequently governments have to request the measurement and use of personal data from students. In some educational settings across the EU students are already being asked to agree to have data tracked, often

assuming that their study will benefit by opting in - for example the Open University in the UK routinely asks students permission to record and use their data.

Educational institutions capture student data under the pretext of providing personalisation of learning preferences and adaptive recommender systems. The rationalisation for student tracking is made in statistical or mathematical terms, citing the advantages of getting better insights due to bigger data. But who really benefits from big data and student tracking – governments, institutions, students, or third party organisations?

Ethical implications of learning analytics are likely to become even more complicated if social media data becomes part of student monitoring. Ethical recommendations put forward by the UK Government-funded social media research group include core principles such as informed consent, avoidance of personal and social harm, and non-disclosure of identity (GSR, 2016). However, these recommendations in themselves are not sufficient to protect the right to privacy of individual students. When data from multiple, disjointed data-sets are brought together, it becomes easier to decipher individual identities, even if masking techniques are deployed. For example, a student may post an anonymous message in a social media channel, but when this data is combined with school data, geo-location data, or even a different social media platform, it may reveal the identity of the student.

Ethical considerations are not limited to data collection or analysis, but also extend to data storage and ownership issues. When commercial institutions collect, and own student tracking data, there is a risk of data being transferred to a third party and being used for purposes that are not aligned with the original intention of the data collection. Conventional privacy measures for consent mechanisms lack the transparency needed for educational data (D' Acquisto et al., 2015). This is especially important in the case of student tracking data, as the European Commission's report on Ethics for Researcher's (Commission, 2013a) stipulates that conventional, informed consent procedures are inadequate for children's data.

Large quantities of personal data are already being gathered sometimes without the knowledge or consent of each individual student. The EU, individual nation governments and policy makers have an important role to play beyond data policing and data governance to include protecting the civil rights of EU citizens and encouraging their empowerment through the current digital transformation. This includes ensuring every citizen develops a 'digital mindedness' and understanding not only of *how* to use the internet, but *why* it is important to be fully engaged in online activity. This ability to understand the role of the internet in our everyday lives is critical for digital inclusion.

Implications for EU policy

While this section has indicated the potential for big data in tracking students, the actual benefits of tracking with big data have yet to be substantially demonstrated. A key question to be considered by EU policy makers is whether student tracking with big data is the right direction to take European education. Important trends and questions highlighted in the previous sections point to the following key issues and observations for policy attention:

Does student tracking promote equity? If big data is to be used to address the inequality between poor and high achieving students, whose performance is often influenced by environmental factors, the data has to extend beyond current narrow measures of 'success' in education (e.g. completion rates, grades etc.). A wider range of data on student performance could be used to implement and evaluate educational policies on a national level.

Can algorithms track student progress objectively? The problem of algorithm bias is one of the most significant challenges in student tracking. Attempts to mitigate this

problem include co-design methodologies that bring coders into dialogue with the users of the algorithms. However, while these methods help take some factors into consideration, they do not integrate the range of different types of expertise needed to for unbiased algorithms. Nor does it allow for the limitations of the data.

Also people who have limited opportunity to contribute to big datasets essentially are excluded from digital society. This means that, by not gathering data from everyone in society, algorithms indirectly are contributing to the digital divide. To reduce the biases inherent in algorithms, quantitatively-data driven decisions should be supplemented with more complex qualitative data. It is also important to ensure everyone understands the influence of algorithms on society, just as we need people to understand the bias of media on society (Leu et al., 2013, Voogt et al., 2013, Littlejohn et al., 2012).

Can student tracking with big data be inclusive? If student tracking with big data is to be implemented across the EU, it is critical to find ways to ensure everyone is included in data monitoring, not only those who already hold the privilege of being online. If people from parts of society are not involved in data production, then the data does not take into account the needs of everyone. A potential consequence could be that big data used for education could widen gaps between those who have access to technology and those who do not.

Can social data be used in educational tracking? Even where it is possible to use data generated through online activity to aid student tracking, this data is not always the right data needed to monitor and support teaching and learning. Transparency of data will be an important issue, particularly in terms of who has access to what data, and to what extent each individual can choose what they share, and with who. Therefore, it is important that EU policies consider the implications of using social data for educational tracking.

How can governments deal with issues of data ownership? As the ownership of data becomes blurred, policies must be reviewed in response to this changing educational digital landscape to consider the use of data generated both across and outside educational institutions.

How should ethics be framed in terms of big data? Ethical considerations with respect to student tracking are complex. People and organisations need to use data already available in an ethical way through the development of basic skills not just on *how* to use data but *wh*y. There has already been progress in terms of framing ethical considerations, but this work is in its infancy and will require a variety of expertise. More consideration has to be given to developing ethical guidelines for student tracking.

Looking ahead

With technological advances in the areas such as learning platforms, big data, and analytics, many of the logistical issues concerned with student tracking could be solved through increasingly refined approaches. The variety and volume of data collected will potentially be able to create sophisticated overviews of how students are performing, and the best ways to guide them on their developmental journey. However, a fundamental question remains as to whether student tracking will be the best way to use big data. The role tracking will play in the next 30 years remains unclear due to the lack of convincing evidence of its benefits in its current form.

In addition, the way that people learn is constantly changing. 30 years ago, it would have been impossible to predict the ubiquitous nature of the Internet and its accompanying educational resources, which have enabled millions to develop new skills informally. Unless governments think about education differently, keeping one step ahead of how people actually learn, they will not be able to capitalise on the potential of big data. Given the way that society's relationship with learning is advancing, tracking in the way that it is currently defined and implemented may not be relevant in the future.

However, if tracking moves its focus from ability streaming, to bringing people with similar interests together, it could become very influential in educational policies and practices. For example, tracking systems utilising big data could become 'recommender systems' for creating groups during project based assignments, Tracking systems could become the foundations of more flexible educational structures that allow people to follow their interests in formal education, in their professional lives and in their homes.

Based on some of the most successful and innovative education systems, such as the new phenomena-based curriculum of Finland[138], in 30 years there could be a comprehensive move away from traditional classrooms to education being based around project work. Education could be significantly more flexible, supporting entrepreneurs who choose to learn by creating their own business at a young age, in addition to more traditional career paths such as doctors and lawyers, many of whose professions will change fundamentally through the introduction of robotics and artificial intelligence. Education has to become more flexible to adapt to these changes in society.

There is a set of issues around tracking that must be tackled if systems based on big data will become the basis for improving student education:

- **The effectiveness of tracking.** There is evidence both for and against the kinds of tracking systems employed by countries currently, but most research questions its effectiveness. This could be the ideal time for governments to re-evaluate the tracking process, using it to bring together people who will complement each other well, rather than solely those who perform well academically;
- **Privacy**. In any tracking system making use of big data privacy must be a priority from the beginning. Thinking about privacy from the outset is one factor that enabled Estonia to build a tracking system in which people have trust. However, trust in governments varies across the EU, therefore there could be difficulties in introducing systems in countries with more complexity than Estonia (e.g., diversity in population, size of country or culture). Privacy (as emphasised in the privacy section of this study) must therefore be taken into consideration from the outset, including data transparency. These are difficult problems, as even data scientists are not currently clear about issues around the transparency of data;
- **Bias of algorithms**. The algorithms needed to attain a future where tracking allows flexible choices will need to be complex. It is very possible in that situation that even the people writing the algorithms will not have a clear view of the bias encoded in the algorithms. This problem can be overcome by taking an interdisciplinary approach to writing algorithms using methods where people with different types of knowledge (for example, computer science, information science, learning science, sociology, anthropology) work together to 'co-design' new algorithms. While co-design processes might reduce bias, it should be recognised that such bias will always exist. However, it can be argued that bias is an integral aspect of education already, for example where teachers have to make subjective judgements about a student's progress;
- **Data gathering**. As technology advances gathering data is likely to be more streamlined than before. However, the issue to consider will be is the right data being gathered? Governments will need to rethink their educational models, and the role that tracking can play in them as fundamental changes in education systems will likely be needed. If policies are enforced that underpin education with systems based on conventional educational models, an opportunity will be missed for future development and change;

[138] http://www.oph.fi/english/current_issues/101/0/subject_teaching_in_finnish_schools_is_not_being_abolished

- **Ethics**. If tracking is implemented in its current form, many ethical questions will need to be considered. For example, should there be an element of choice for learners who may, or may not, want some of their data to be used? Trust underpins ethics.

Big data has the potential to help solve some of these issues, such as reducing inequality due to biased selection processes in tracking. However, many of the issues outlined above require assessment of underlying assumptions about education. Before investing funding into tracking, governments must consider fundamental questions about the effectiveness of tracking and where tracking will fit within evolving educational practice across the EU.

Conclusions

The use of tracking systems has been observed in European educational practices for decades, often falling in and out of fashion due to political agendas (Boaler, 1997). At this stage, it is too early to confirm whether big data and student tracking has the potential to transform education policy and practice, addressing the largest criticism of student tracking: worsening inequality between high and low achieving students. The complexity associated with the use of big data is used in education is only just being realised.

There are many questions about the objectivity and ethics of using big data to make a decision that will potentially change the life path of every student in a country. Coupled with these are issues with the basis for tracking being a beneficial educational model. While there are some countries who seem to succeed using a tracking system research has provided scare evidence for its effectiveness.

Advancing the use of big data in education requires insight through multi-disciplinary expertise from a unique mix of social science and ethics experts along with technology specialists. Rather than trying to improve weaknesses in existing systems technology could hold the potential for completely new ways of providing individually tailored educational experiences.

Skills forecasting

Introduction and Context

This section explores the advantages and challenges of using big data analysis in the process of aligning the skills demand on the labour market with the skills supply provided by educational establishments at all education levels. This includes skills development tailored to the labour market through VET, as well as the overall coherence across the whole educational system to develop (and monitor) skills.

First, the section identifies ways in which big data can be implemented in the analysis of labour market demands. Second, it outlines possible avenues of using educational big data to help developing students' skills and to improve the responsiveness of educational systems to labour market skills demand. Finally, the opportunities and challenges that are discussed are taken into possible actions at EU level with the purpose of creating a framework for implementation of big data in education, with a view to increase the relevance of skills supply for the European labour market.

The Europe 2020 strategy[139] sets the agenda for smart, sustainable and inclusive economic growth through knowledge and innovation. This strategy defines employment objectives (a target of 75% for 2020), as well as quality education and the development of a skilled workforce across the EU. This strategy is geared towards a knowledge-based economy competitive in a global and digital world. Policies such as the 2009 strategic framework (Commission, 2009) for European Cooperation in Education and Training (ET 2020), and its 2015 implementation report (Commission, 2015) have highlighted the need for effectiveness and efficiency in raising the skills and competences of the European workforce.

In its Communication on a new and comprehensive Skills Agenda for Europe, the Commission identifies *skills mismatch* as a key challenge (Commission, 2016c, Commission, 2016d). Despite a rather large availability of labour, European employers are often dissatisfied with applicants' skills. In 2013 27% of employers had reported that they have left a vacancy open in the past year because they could not find anyone with the right skills, and 33% said the lack of skills is causing major business problems in the form of cost, quality or time (Barton et al., 2013).

Due to the dynamic and asymmetry of information that characterise labour markets, different types of skills mismatches coexist:

1. **Skill shortage** occurs when the demand for a particular type of skill exceeds the supply of people that skill at equilibrium rates of pay;
2. **Qualification mismatch** occurs when the level of qualification is different from that required to perform the job adequately which generates:
3. **Over or under-qualification** arising when the level of qualification or education is higher or lower than required to perform the job adequately;
4. **Skill gaps** which reflect the difference between the type and level of skills that a candidate possesses and the one required to perform the job adequately;
5. **Over or under-skilling** arises when the level of skill is higher or lower than required to perform the job adequately. (WEF, 2014)

Matching skills requirements on the labour market with job seekers' skills is a key priority for the European Union, for example through the work of the European Centre for the Development of Vocational Training (Cedefop), which analyses skills supply and demand

[139] http://ec.europa.eu/europe2020/europe-2020-in-a-nutshell/index_en.htm

at EU level. In 2014, Cedefop carried out a first European skills and jobs survey to provide empirical evidence on existing skills mismatches across Europe.

This paper analyses how big data can contribute to tackling skills mismatches and support skills forecast, but it should be noted that dealing with the issue of skills mismatches requires a broader approach, and entails a comprehensive long-term strategy involving public-private partnership between governments, employers, unions and education institutions and systems.

In recent years, ICT-related innovations have created new forms and types of data 'big data', that can be collected and interpreted with the purpose of efficiently and comprehensively forecasting skills on the labour market. Furthermore, advances in technology provide solutions to align skills forecasting with educational programmes, in order to better prepare pupils and students for the workforce. On the other hand, embedding big data analysis in skills development across educational pathways can help increase pupils' and students' academic performances and help them make personalised choices in careers to follow and jobs to start with in the workforce.

The central advantage of big data is that large amounts of real-time data can be continuously acquired, thus making analysis more timely, more individualised, and potentially less inexpensive (WEF, 2014). Moreover, big data reflect better the behaviour of individuals in the context of their life, which avoids biases like social desirability in responses to classical surveys. Through fine-grained analysis of big data, decision-makers can be supported in their decision-making process. So far, big data has been mostly used in the fields of banking, insurances, health, energy, transport and IT and only a few approaches can be found in the labour market monitoring (Askitas and Zimmermann, 2015). In the context of skills forecasting, big data is generated at two levels:

- In the labour market, where big data is generated across online job postings and social media (notably LinkedIn but also other social media, as explained in the following section);
- Across educational systems, where ICT uptake is starting to generate a digital footprint of students through software and teaching and learning platforms that are used in areas such as school management, course delivery, and for assessment.

Big data in the context of skills forecasting is very much an emerging practice. Digital equipment across educational structures is unevenly developed in Europe, especially in primary and secondary schools, as shown by a 2013 study for the European Commission (EUN, 2013). This means that limited amounts of big data have been generated on skills supply, and even less has been processed and analysed.

In keeping with the objectives of the above-mentioned policy documents, the objective of this section is to discuss the advantages and challenges of using developments in big data analysis to tackle skills mismatch on the European labour market and the implications for educational systems and their monitoring. The approach is focused on two steps within the larger 'matching skills' process, where the use of big data can be best integrated:

- Monitoring labour markets: Forecasting of skill needs on the labour markets;
- Monitoring and governance of educational systems: Improving the responsiveness of education to the identified labour market needs.

For the purpose of this analysis, the section makes use of previous studies of the labour market that employ big data and identifies useful findings in relation to skills forecasting, and also considers challenges in big data collection and analysis at EU level.

It further investigates current examples of the use of big data in education that provide ways in which big data analysis helps align skills demand and supply. It also identifies the challenges in implementing big data infrastructures in EU educational systems and the related governance issues.

Key issues and Challenges

The section first explores ways in which big data analysis can help monitoring skills requirements on labour markets. Second, it focuses on the skills supply side, by discussing ways in which the use of big data could or is already integrated in all-level educational programmes in order to monitor the development of pupils' and students' skills. It addresses the issues related to aligning skills demand and supply, and shows the potential for big data analysis to tackle these issues.

Using Big Data in monitoring skills demands on the labour market

A relevant context in the application of big data in labour market is through online job portals, as they provide a growing quantity of information on the demand and supply. There is an opportunity to capture the needs of employers according to sectors, professions and skills. Also, job portals allow a complex insight into the situation of applicants, as their CVs (if data are robust and in clear structure) can be systematically analysed. Where they are geographically referenced it can be possible to analyse at spatial levels (e.g. at regional, national, or EU-level).

Also, social media is a growing source of big data on analysing the labour market. Social media platforms are increasingly becoming vehicles for innovative and effective services, among which employment is highly represented. For instance, social recruiting is a new form of matching the supply and the demand of labour on the web, made possible in the context of growing opportunities for building relations and facilitating communication via social networks. In January 2017, it was reported that "*Deutsche Bank launched a programme late last year to monitor the online activity of university students to identify those who might be a good fit for the bank but would not apply through traditional channels such as on-campus recruitment drive*" (Noonan, 2017).

In 2014, 61% of the global recruitment activity involved the use of the Internet, and the use of LinkedIn was predominant (68%) in assessing a candidate's reputation and finding job opportunities (ADECCO, 2014). Facebook is also used as a data source in a lesser extent. The most relevant data sets which can be retrieved from social networks for recruitment purposes express 'tangible skills' reflected in previous work experiences (attractiveness index of 0.63 out of 1 for recruiters), followed by the presence of professional prizes or awards (0.38) and the personality insights that can be identified from the profile (0.32) (ADECCO, 2014).

Working with big data rather than traditional surveys in order to collect information on skills demand has some clear advantages:

- **Data scraping:** the cost of collecting big data is lower than the cost of traditional data collection methods;
- **Time to market:** data are much more up-to-date enabling the use of real-time analysis techniques;
- **Bottom-up approach:** the raw data collected online do not emerge from pre-defined taxonomies like it happens in the case of classic surveys, therefore data are richer and closest to reality. This is extremely useful when identifying personal and professional skills which are expressed freely during online activities and do not fit in pre-defined option answers (Larsen et al., 2015).

Until now, a limited number of studies on skills demand have been concluded using big data at regional and national levels. The majority focused on online portals that publish job vacancies (on Job Boards) (Kureková et al., 2014).

Capiluppi and Baravalle investigated the skills demand for the IT personnel in the UK and to what extent the needed skills were delivered by universities (Capiluppi and Baravalle, 2010). The authors developed a 'web spider' to scrape vacancies from the leading private internet recruitment site monster.com,[140] and then analysed the skills required in comparison with the skills provided. The research revealed that the request for specific IT skills has been constant in the selected time frame (September 2009 – May 2010), but that job adverts tend not to be very specific in stating the details for needed technical skills (i.e. demanded versions of operating systems, specific commands for programming languages etc.).

Štefánik studied online data from a private job portal in Slovakia, analysing both vacancies and CV data (Štefánik, 2012). He concentrated on the labour market segment of the highly skilled and examined the matching of demand and supply of university graduates for a limited number of narrowly defined highly skilled professions. The findings show that the representativeness of Internet job search data is limited primarily by the penetration of Internet usage. Within the occupational structure there is an overrepresentation of clerks and service workers, and a slight underrepresentation of managers and professionals. Within the economic sector structure, there is an overrepresentation of private services and an underrepresentation of public services.

Dusi, Mercorio and Mezzanzanica analysed the labour market dynamics in Italy focusing on web job vacancies with the purpose to reach a matching between the occupations sought in the market and the corresponding skills (Dusi et al., 2015). The authors built a portal called WollyBi[141] which aims to support the activity of employment agencies, public employment services, trade unions and VET actors through an in-depth analysis of labour market demands focusing on three dimensions: territory, professions, skills (Larsen et al., 2015).

At EU level, the Cedefop is currently undertaking research to detect emerging skill needs in European labour markets with the assistance of big data. The researchers developed a new web-scraping tool of job market vacancies which focuses on extracting information on skills and job requirements in five EU countries.[142]

Key challenges identified in using big data to analyse labour market skills demand

The experiences of research into online vacancy data in the labour market has revealed some consistent challenges.

Sample Representativeness. When working with web job portals, a difficulty lies in assessing whether the sample of online job vacancies is representative of all job vacancies in a specified labour market. Not all vacancies are advertised online, much hiring takes place internally or through informal means and networks, especially at local level. Therefore, there is a share of vacancies which will fall outside the population sample.

Given this limit, using big data in monitoring labour markets must be combined with existing information from traditional statistical surveys on the occupations and skills sets that are most needed at a given moment, on the required level of those skills and their

[140] http://www.monster.co.uk/geo/siteselection
[141] http://www.wollybi.com/
[142] http://www.cedefop.europa.eu/en/news-and-press/news/cedefops-research-big-data-and-skill-mismatch-global-knowledge-frontier

existing level in the workforce.[143] Interviews with recruiters and HR managers are complementary.

Data distribution across education levels. Research shows that the web is mainly a channel for hiring high educated professionals (WEF, 2014). Low skilled job positions related to compulsory schooling are almost absent from the web portals. This might lead to a biased assessment of skills needs, by emphasising the need for higher education graduates at the expense of lower educated individuals.

Methodological issues. Occupational titles can often be ambiguous and fail to reflect the true nature of the skills required in the vacancy. This poses a challenge in sorting vacancies into occupational groups based on job titles. While occupational title is the essential starting point, more details are necessary to better understand skills requirements. Therefore, powerful algorithms need to be developed in order to extract meaningful information about demanded skills from unstructured data like job descriptions, which is commonly disregarded in classic skills analysis (Wowczko, 2015).

Using Big Data in monitoring educational systems in relations to skills development

Education produces significant volumes of data because it involves hours of both individual and group work (in classes and at home) for all educational years (from primary to higher education). This translates into complex interactions between students and didactic materials (training manuals, books, quizzes etc.). These interactions and their results (academic projects, dissertations, simulations, films etc.) contain a wealth of useful information on pupils' and students' learning and performance. Advances in technology and data science make it possible to:

- Continuously record all these interactions and results, through the use by teachers and students of an increasing number of digital devices and tools in the classroom and at home like: computers, tablets, applications, MOOCs, online repositories etc.;
- Collect and centralise all this educational data through specially developed techniques such as data scraping, and data crawling;
- Process and analyse educational data via dedicated tools (analytics) and software frameworks.[144]

Knewton[145] (a US-based company specialised in educational data processing and adaptive learning) divides educational data into five types: one stemming from student identity, and four from student activity-based data sets that have the potential to improve learning outcomes (Ferreira, 2015).

- **Identity Data** pertain to student identity, rights of use of different didactic applications and demographic information (age, gender, race, nationality, income);
- **User Interaction Data** refer to student behaviour when in contact with different applications, websites, online repositories etc. It includes engagement metrics, click rate, page views, bounce rate and is the easiest to collect of the data sets;

[143] Including notably the EU Labour Force Survey (EU-LFS), The OECD Programme for International Student Assessment (PISA), The OECD Survey of Adult Skills (PIAAC), The European Working Conditions Survey (EWCS), Cedefop's European Skills and Jobs (ESJ) survey, The EU Adult Education Survey (AES)The EU Continuing Vocational Training Survey (CVTS)

[144] A software framework is a universal, reusable software environment that provides particular functionalities to facilitate development of software applications, products and solutions (Wikipedia). In relation to Big Data analysis, the most popular software framework is Hadoop, which is used to store massive amounts of data and running applications to process that data.

[145] https://www.knewton.com/

- **Inferred Content Data** reflect the efficacy of instructional materials in student proficiency gains (of students using a certain pool of resources), such as what measurable proficiency gains result when a student interacts with a certain piece of content, or how well does a question actually assess what it intends to?
- **System-Wide Data** refer to grades, disciplinary records, attendance information etc. This type of data become useful at very large scale as it may bring out inferences and correlations which can help shape recommendations at educational system level (e.g. correlation between the attendance rate and the structure of classes);
- **Inferred Student Data** reveal what concepts a student knows and at what percentage of proficiency. It helps predicting if the proficiency level will grow or decrease in time. This type of information can be inferred from quiz results, syllabus browsing, behaviour when using an app etc. Inferred student data are instrumental in brushing the palette of skills a student has developed or will likely develop at a certain schooling level.

Inferred student data are the most difficult to generate. This means that the 'digital footprint' of students is much more difficult to gather and analyse when it comes to skills. Developing a system capable of gathering and analysing this type of data requires specific teams of teachers, course designers, technologists and data scientists, a big amount of content and a large number of students and instructors interacting with that content. It should also be noted that this approach largely ignores the social and cultural roles of educational systems.[146]

Educational big data processing, and analysis, provides insights which are crucial in improving pupils' and students' learning and performance, teachers' performance, administration's efficiency and also educational programmes' relevance for the labour market.

Using Big Data in aligning skills demand and supply

The forecast of skills demand on the labour market should impact educational programmes in order to ensure that pupils and students are taught the right skills to find the right job. Big data are also seen as a way to increase the efficiency of skills acquisition and increase the responsiveness of educational structures to the learning curves of students (van Rijmenam, 2014).

Advances in big data analytics have the potential to be very effective in the process of consolidating the communication between the workforce and all levels of educational systems, by making available for the education providers a series of metrics on the needs of a company in terms of sectorial occupations, expertise or performance. Such an approach is being implemented to this purpose in the USA. Currently, 43 states link K-12 (primary and secondary schools – equivalent to ISCED 1-3) with post-secondary data systems, 19 states link K-12 and workforce systems, and 27 states connect post-secondary and workforce data systems.

The sharing of important data about graduate performances at work has empowered universities in Pennsylvania for instance in the last five years to develop several new degree programs, discontinue others and work with industry partners to develop special programmes that cultivate needed skills in different sectors (Reid-Martinez and Mathews, 2015).

[146] For example, if we consider the four pillars of education: 1) Learning to know; 2) Learning to do; 3) Learning to be; and 4) Learning to live together. Big data for skills forecasting clearly focuses on the first two. UNESCO. 1996. *Learning: The Treasure Within*. UNESCO. Available: http://unesdoc.unesco.org/images/0010/001095/109590eo.pdf. [Accessed February 28 2017].

However, there is limited knowledge on exactly which indicators in the work environment are clearly determined by the employee skills level. For example, some universities in the United States are using employee *work performance* records in order to assess their skills levels and thus rethink educational programmes that develop the required levels of skills in order for future graduates to succeed in the workface, but to our knowledge, little research has been undertaken on the connection between skills level and work performance.

Research suggests that, for technical occupations like webmasters, soft skills like organisational skills could have a bigger impact on work performance than technical skills specific to the mentioned job (Wade and Parent, 2002). As such, further studies need to be conducted (at a macro-level and on several professions) in order to clearly establish where exactly acquired skills do intervene in the general work experience. This information will guide data scientists in monitoring the right metrics in businesses' activity which will help in correctly assessing what kind of skills (and at which level) are needed to be developed in educational programmes in order to improve the integration of graduates in the workforce.

At higher education level, the Oral Roberts University in Oklahoma[147] is developing applications which help students predictively analyse themselves and gain greater understanding of what they need to do to start a career and find a job. The apps contain also a 'My Life Data'[148] button which compiles data regarding students' educational experience. Typically, this includes student's engagement level, use of classroom resources, academic performance, attention span, language and vocabulary use. Based on this data, the buttons give milestones, such as 'You've reached 85% of your academic goal' or 'Here are the jobs that are available according to the credentials that you acquired', or 'You are rated within the top 15% of students who could apply for specific jobs' (Reid-Martinez and Mathews, 2015).

In order to further ease the transitions from university to work, apprenticeships and work-based learning are gaining ground. A growing number of companies provide training to young graduates to ensure they will develop the right skills at the right level of proficiency a job requires. However, in the age of international business and global competition, organisations struggle to identify the training topics that provide the best returns on investments. Big data analytics can reveal more accurate training needs. Training managers can use big data to learn about and fill the gaps in organisational skills and knowledge and select the best methods to teach trainees at all levels.

Using big data to improve business training also helps employers gain insight into the most effective curricula, platforms and tools in helping trainees develop the right skills. Big data can also differentiate among people in various roles with various backgrounds, triggering personalised approaches to training (Phillips, 2016). Insights that employers develop into effective curricula and personalised learning of skills can be successfully used by education providers to improve educational programmes for students. This type of collaboration is integrated into the highly emergent dual training systems which, in addition to providing work-based learning, feature cooperation between employers and public authorities to govern education and provide the integration of theory and practice through cooperation between education providers and employers in skills development (WEF, 2014).

There is little data to show how skills acquired at lower educational levels (as secondary school or vocational training) can be aligned with skills demand for appropriate jobs. More research should be undertaken to establish how and what skills acquired at each level of education can build into different employee profiles which are of interest for

[147] www.oru.edu/
[148] See 'Whole Person Assessment' at http://handbook.oru.edu/section-3/#student_services

employers. This knowledge would help data scientists collect more meaningful big data from each educational level and analyse it in relation to specific requirements from the labour market, in order to assess the efficiency of educational programmes in aligning with market needs.

Analysing big data at each educational level with this purpose would lead to emerging evolutionary patterns throughout education with direct relevance for employers. Implementing big data analysis at university level is still a nascent field (Burns, 2016a). At K-12 levels, the purpose of big data analysis is limited to assessing pupils' performance in relation to the established educational programme. Monitoring acquired skills and their degree of proficiency is mostly during higher education, as explained in the above-mentioned examples, whereas a meaningful educational big data analysis of acquired skills necessary in the labour market requires a more consistent approach across the different educational levels.

Developing big data infrastructure, and analytics solutions to connect the workforce needs with educational programmes nonetheless requires complex and costly big data techniques and analysis tools, which most education providers do not possess or are incapable of building functionalities where these techniques can be applied. This is why schools and even universities are mostly using solutions developed in the private sector for the provision of big data techniques. As such, there is a growing market for suppliers of commercial technology in the education technology (EdTech) industry. These suppliers are offering their professional services for the mining, collecting, processing and analysing of student data, and even for applying data-driven decision-making processes (Har Carmel, 2016).

Corporations such as Pearson, McGraw-Hill, Knewton or Khan Academy, are marketing a wide variety of big data-based technologies to educational establishments. The technologies can be applied in all aspects of digital education allowing the implementation of personalised learning, adaptive learning, accurate assessment, effective feedback or performance prediction.

For instance, based on the analysis of data driven from pupils' test results and assessments, Knewton is able to assess what each student in a class knows and what they are struggling with. Given this information and the goals they are working towards, Knewton developed individual adaptive programmes for each student to work on in real time. Adaptive learning is fuelling blended learning programmes, which focuses on teaching different skillsets to students through carefully curated software, based on student-set goals. In an hour of blended learning, students can choose to practice math skills, read a book, use online grammar or spelling software, practice typing skills, or work in any other area for which they have a digital on- or offline tool available.[149]

The education technology (EdTech) industry is growing rapidly, as distribution and platforms scale internationally, the market is projected to grow at 17% per annum, reaching $252bn by 2020 (EdTechXGlobal, 2016). In 2015, the global market was estimated to be $43.27bn,[150] and to reach $93.76bn by 2020. To date, the US has set the trend and pace of the EdTech market. Asia is experiencing the world's fastest growth in investment into the sector and Europe has also seen increases in investments and acquisitions. However, Europe remains a largely under-invested EdTech market.

The EdTech market is clearly a high-potential and fast-growing market, but is not yet mature. An increasing number of start-ups are developing in the sector, which generates a fragmented approach on exploring big data in education. In the US, the way educational data is being set up, housed, maintained and governed is different in every

[149] https://www.knewton.com/resources/blog/ed-tech/blended-learning/
[150] http://www.marketsandmarkets.com/PressReleases/educational-technology-ed-tech.asp

school district, depending on the local policies and also on the methodologies and capacities each tech company deploys. Big data needs to be analysed at macro-level, equally exploring the entire ecosystem of the school district and universities, in order to produce more meaningful results (Harven, 2013), in the way mentioned above.

Another concern – especially when considering EU policy-making implications – is equity in education. Access to EdTech solutions is highly conditioned by access to and quality of digital tools available to students. By introducing several tools facilitating out-of-classroom learning, particular attention should be paid to potential social discriminations; whilst an ever larger proportion of households have access to internet and a personal computer, quality of this access is highly heterogeneous in terms of internet speed, number and quality of devices available, level of digital literacy among the household. Taking content out of the classroom also means that households play a more important role as a learning place. While edTech solutions can improve the quality of digital learning materials, little can they do about the learning environment of students at home and discriminations stemming from disparities across social background (Watters, 2015).

Digital readiness in education for data generation and collection

The potential volume and diversity of educational big data is directly dependent on the level of ICT infrastructure in educational establishments, which still needs to be improved at EU level. According to a 2013 EU survey, school heads and teachers consider that insufficient ICT equipment (especially interactive white boards, laptops and PCs) is the major obstacle to ICT use (Wastiau et al., 2013). There are between 3 and 7 students per computer on average in Europe. Also, 37% of grade 4, 24% of grade 8, 55% of grade 11 general and 50% of grade 11 vocational students are in highly digitally equipped schools throughout Europe (EUN, 2013).

On the other hand, generating and collecting quality data sets on student experiences and performance, requires that educational establishments deploy teams of teachers and instructors (and managers) well-versed in the manipulation of digital assets (computers, tablets, software, online apps), capable of showing students how to interact with digital learning content. As such, educators' digital competences and the frequency of ICT-based learning activities are instrumental in laying out a big data infrastructure in the educational system.

Overcoming fragmentation of data across educational levels for meaningful data analysis

Given the recent expansion in the EdTech market, the approaches in handling big data in educations are immature and heterogeneous. Educational big data collection and analysis has not been given a consistent purpose in relation to skills formation at all educational levels. Furthermore, applying educational big data analytics differently at each educational institution (depending on the contracted company) is not useful in generating a meaningful picture on what skills a student is developing through school and in what way. Analysis is needed in the same way at regional level (on as many education providers in the region as possible), with the same purpose and with a possibility of joint data analysis at national (or European level).

This nonetheless implies some degree of harmonisation in infrastructure, tools and practices in big data analytics to ensure comparability of data and the relevance of results. The current trend towards outsourcing of IT also raises potential issues on the interoperability of data across service providers.

Integration of data processing in educational structures

Data processing solutions for education are extremely difficult to develop and to understand, given the high level of required expertise in domains like computer science, statistics, mathematics and analytics. There is a challenge to build comprehensive

solutions which integrate both straightforward data (like user interaction data) and more-difficult-to-attain data sets (like skills-related data such as inferred student data or inferred content data) in order to deliver most efficient and realistic results which could improve educational systems responsiveness to both labour market and student demands. This implies that the organisations building these solutions make them as easy to integrate as possible in the educational framework, so that institutions can get the most valuable information.

The technical systems for conducting educational big data collection and analysis are rather expensive to build[151]. Implementing such solutions require cooperation between schools, with education ministries and statistical offices to alleviate the burden on individual structures: for example, by developing/making available simple data processing tools and provide data analysis services at national level.

Protecting student privacy

Contracting third-party actors to improve education processes by use of big data raised concerns across the USA about a possible misuse of student data and breaches in student privacy. Critics have also been concerned that monitoring student activities may limit creativity, free speech and free thought, by creating a *surveillance effect* (Zeide, 2016). In this context, regulatory reforms to protect student data have taken place in the USA. The EU data protection law may need to be updated so it could specifically deal with the possible dangers of big data in education. These issues are developed specifically in another section in this document.

Further conceptual and practical issues in aligning skills demand and supply

Even though educational and occupational proxies are the two main reference points in studying the skills demand and supply on the labour market, there is no widely accepted and available standard classification for job skills requirements across countries comparable with the International Standard Classification of Occupations (ISCO) or the International Standard Classification of Education (ISCED) (Handel, 2012).

Even at country level, standardised information on job tasks for the national workforce is almost non-existent. This makes research on skills requirements at national level (and even more so at EU level) very difficult, as researchers are forced to rely on indirect measures of jobs skills requirements, like job titles, which are rather nominal values that offer little information on skills. Additionally, the rigidity of a universal framework of qualifications imposes another difficulty; the rapid technological advance adds different qualifications to a job which are not foreseen in the existing classification.

Another measure of skills is education level, based on the education level of the workers currently active in each occupation. However educational level is often used as a credential to regulate access to jobs on the basis on social and cultural capital, rather than serving as functional requirement. This implies that education is not a relevant measure of skills demand, which deepens the gap between skills demand and the competencies acquired in schools and universities.

Data mining and data analysis would support the European Commission's policies (such as the new skills agenda for Europe) aimed at improving the teaching and recognition of skills by creating real-time views on skills demand on the labour market. In the long run, a comparison based on skills requirements taken at different moments in time would improve information on the evolution of skills demand. This will help in refining existing qualifications framework (such as the European Reference Frameworks for Key Competences) based on actual skills requirements, and equip the education system with

[151] https://codeactsineducation.wordpress.com/2016/06/02/critical-questions-for-big-data-in-education/

tools to better anticipate skill needs. Other aspects need to be taken on board for such frameworks: skills requirements based on big data need to be combined with more holistic skills developments, such as "*critical thinking, independent judgement, problem-solving, and information and media literacy skills*" (UNESCO, 2015). Whilst many of these skills will not emerge directly from the labour market, they are a key aspect of a humanistic and integrated approach to educational systems.

Minimising the incongruences between skills demand and supply is a long-term objective, like education designed as a long-term process (which goes beyond the simple acquisition of skills). Big data analysis could successfully align the highly dynamic labour environment (due to rapid technologic evolution) with slower-paced education system by being able to offer, in time, predictive models on the evolution of skills demand which could be integrated in relevant lifelong learning programmes. Conversely, the desirability of an educational system linked to labour market evolutions and employability is questionable. By developing its responsiveness and capacity to address existing skills gaps, educational systems should not be encouraged to overlook the broader role and purposes of education (active citizenship, social inclusion and awareness, education equity, etc…) and go "beyond narrow utilitarianism and economism to integrate the multiple dimensions of human existence" (UNESCO, 2015). Even from the perspective of employability, the capacity to develop new and unforeseen skillsets also remains highly relevant, whereas skills forecasting thus far remains largely limited to predicting the evolution of existing classifications. In other words, the use of big data for skills forecasting will be helpful to alleviate issues linked to skills mismatches across existing business models. Its effectiveness to anticipate the skills required to create new business models (and related jobs & skills) is however highly uncertain.

Implications for EU policy

Using big data analysis for skill forecasting is relatively new. Its uptake is underway, but it is not expected to reach its full potential before the medium term (8-10 years), and initially in Member States where digitisation of schools and Higher Education structures is higher and with dedicated structures to analyse big data. Examples include the UK Higher Education Statistics Agency (HESA) work on data futures,[152] or in Denmark where the newly-launched Danish Center for Big Data Analytics driven Innovation (DABAI) includes educational big data as a priority[153]. In the short term, policy measures could encourage the generation of exploitable big data in specific sectors, with priority areas being:

- To facilitate the creation of a robust infrastructure and methodological framework enabling data collection and data analysis;
- To establish a regulatory framework that takes into account data privacy concerns and governance of big data for education.

As such, we highlight some potential EU-level policy measures to deal with the current challenges identified in this paper. Our recommendations focus on: 1) labour market big data analysis; 2) educational big data analysis; 3) governance of educational systems; and, 4) big data framework implementation in schools and higher education.

Labour market big data analysis

Work with local bodies and social partners (e.g. through the ET 2020 working group mechanism) in order to encourage local job agencies to develop online job portals relevant at local level to cover a wider spectrum of job vacancies, which are not typically advertised online (such as those involving lower level skills which circulate in more informal local networks). Market coverage and technical advancement of job portals

[152] https://www.hesa.ac.uk/innovation/data-futures
[153] http://dabai.dk/en/research-domains/domain-educational-data

differ at the EU level, which lowers the quality of data collected and increases granularity. In particular, local skills requirements are not always reflected in the job vacancy data base of larger job agencies which operate at EU level.

Encourage the digitisation of the job vacancies which involve lower qualifications and vocational skills, to ensure universal access to online job markets regardless of the level of education - as part of proposal for the New Skills Agenda for Europe.

Integrate big data analysis as part of the Cedefop and Eurostat work on skills forecasting to supplement existing surveys: such as, the European skills and jobs survey, and ESSnet[154] work on big data and skills, and analyses. For instance, web scraping techniques can help to refine data on job occupations and qualifications by going beyond standardised categories to assess emerging trends in demand for skills and competences. In the longer run, big data analysis should be gradually used on skills supply.

Embedding big data analysis in educational systems

Commission studies to better clarify in which way the skills acquired influence work performance and contribute to employability in order to help data scientists define the metrics which need to be monitored (and thus collect meaningful data) to help improve the sets of skills developed throughout education. Studies should be conducted at macro-level (regional or EU-level) on several professions which integrate different sets of skills.

With a shorter-term perspective, **support pilot projects** analysing the digital footprint of skills development and potential limitations for big data processing (for example via H2020 Marie Skłodowska-Curie actions). Particular attention should be given to data interoperability and the challenges posed by the (potential) granularity of data sources. At this stage, this approach could be tested on existing online courses to obtain a sufficient amount of data.

Monitor emerging data sources ('tomorrow's big data') and embed automated software or delivery tools facilitating big data analysis where relevant: for example, in the upcoming Commission proposal for an initiative on Graduate Tracking, due in 2nd quarter 2017. The rapid development of MOOCs in Europe is also a promising avenue for big data analysis of skills supply, albeit at higher educational level only (and lifelong learning) for now. A more widespread uptake of ICT solutions in primary and secondary education is however very likely in the coming years.

Governance of educational systems

Work with Education Ministries and (through Eurostat) national statistics offices to assess the feasibility of national big data analysis units to support schools, colleges and universities in making sense of the data they collect and process. In parallel, training for educational staff should equip them with the right skills for data processing.

Support initiatives increasing the responsiveness of educational systems to better react to real-time analysis of skill-demand enabled by big data. The above-mentioned MOOCs and more generally CVET can propose adequate solutions provided that they are equipped with adequate capacity for data analysis, processing of information and adaption of courses. This capacity challenge was highlighted in the Staff Working Document of 2015 Joint Report of the Council and the Commission on the implementation of the strategic framework for European cooperation in education and training (ET 2020).

[154] https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/index.php/ESSnet_Big_Data

Assess the potential risks of outsourcing student data and analysis and establish a power balance between data subjects and data users that would protect students' rights. For example, in order to prevent biased data-driven decision-making, a regulatory solution would be to implement auditing systems that review the algorithms and the variables that they use to detect possible discriminating results (Williamson, 2016). Particular attention should be paid to the impact of digital learning on equity in education. For example, increasing the importance of learning from home will deepen disparities across students based on their conditions for studying at home.

Framework for implementation of big data in schools and higher education

Invest financially in capacity building for schools to further support ICT infrastructure for teachers and students. Invest in teachers' professional development in order to increase the number of digitally confident and positive teachers throughout Europe. Effective professional development can transform current positive attitudes in ICT provision into effective digital practice in the classroom.

Support the **creation and dissemination of good quality digital learning resources** to increase the students' interaction with this type of material and therefore increase the amount of usable digital footprint.

Encourage the adaptation of current ICT curricula and educational programmes so that they reflect the evolution of job profiles towards big data professionals and data scientists. Big data skills are becoming increasingly important on the labour market, a trend which is not yet reflected in the supply for these skills: a 2015 study revealed that there are only about 100,000 these highly specialised scientists in Europe and that there is a skill gap of 7.5% of total demand of data workers.[155]

Looking ahead

In the future, more detailed and informative longitudinal analyses of skills requirements will be needed if we are to improve information on the increasingly dynamic developments in skills demand. It is not logistically or economically feasible to expect that all national education and statistical systems can adjust their data collection policies to address future skills needs, but a strategy is needed to provide life-long information related to skills, and to overcome the data discontinuities noted in the skills section.

The section identified that big data analytics can contribute greatly to tackle skills mismatch. This can particularly help in refining existing qualifications framework (such as the European Reference Frameworks for Key Competences) based on actual skills requirements, and then help to equip the education system with knowledge and tools to better anticipate skill needs.

However, aligning skills supply to skills demand does not address future skills needs, so expertise and thought-leadership to anticipate future trends and inform core competences framework is required. An important aspect in this regard is the current lack of big data specialists (and dedicated education and training schemes) to make sense of existing data whilst designing forward-looking analyses. This is however a transversal issue across the different topics covered in this study.

Trends driven by the labour market will arguably decrease the current bias towards more qualified jobs in terms of data for skills demand (linked to the availability of data). This is due to: 1) the shift of skills required, driven by key technological changes (e.g. robotics, nanotechnologies, artificial intelligence, and quantum computing), where low qualifications jobs will become more digital intensive, and where current knowledge-intensive jobs (economists, accountants) will be threatened by intelligent systems; and

[155] https://www.big-data-europe.eu/big-data-labour-market-impacts/

2) automation (Arntz et al., 2016) which will continue to disrupt significant amounts of low-qualified jobs.

In a long-term perspective, a diversification of educational structures is highly likely (digital-enabled learning is virtually ubiquitous and costs of technical solutions very low, so training and learning can be developed by many more organisations – e.g. civil society organisations or companies with niche skills), especially when informed by better evidence, and driven by different types of lifelong learning pathways. This is due to:

- Labour market disruption, as it is expected that in the USA over 30% of men aged 25-54 will not have a job, and that 50% of men should expect to experience unemployment every five years (Summers, 2016);
- Increased retirement age, hence the need to augment skillsets over longer careers, catering for issues such as the need to match jobs not just to skills, but to the physical and cognitive abilities of elderly people;
- A growing role for informal, workplace-based learning beyond more formal training service providers (e.g. through co-working spaces and other forms of shared spaces).

In this evolving context, the role of policy makers and administrations to ensure coherence, as well as to aggregate and analyse data will remain highly relevant. A more partnership–based approach to understand and monitor the contribution of these different structures will be required to overcome potential fragmentation.

Overcoming fragmentation can be undertaken not just by accessing and aggregating sources of big data, but also can stem from individualised skills digital footprints such as e-portfolio of skills, and more intensive use of digital tools at all levels of education (either through reinforced ICT infrastructures or via BYOD). In this scenario, the role of education ministries (and DG EAC) will be instrumental in terms of monitoring and analysis to make sense of individual data and identify trends, and provide relevant feedback to educational structures.

Bibliography

ABDUL-HAMID, J. 2014. *What Matters Most for Education Management Information Systems: A Framework Paper*. World Bank. Published June. Available: http://wbgfiles.worldbank.org/documents/hdn/ed/saber/supporting_doc/Background/EMIS/Framework_SABER-EMIS.pdf. [Accessed November 14 2016].

ADECCO. 2014. *#SocialRecruiting: A Global Study*. ADECCO Group. Available: http://www.adecco.com/en-US/Industry-Insights/Documents/social-recruiting/adecco-global-social-recruiting-survey-global-report.pdf [Accessed June 4 2016].

AERA. 2014. *Standards for Educational and Psychological Testing*. American Educational Research Association. Published July. Available: http://www.aera.net/Publications/Books/Standards-for-Educational-Psychological-Testing-2014-Edition. [Accessed February 16 2017].

ARNETT, J. J. 2000. *Emerging adulthood: A theory of development from the late teens through the twenties*. American Psychologist,, 55, 5, 469-480.

ARNTZ, M., GREGORY, T. & ZIERAHN, U. 2016. *The Risk of Automation for Jobs in OECD Countries: A Comparative Analysis*. OECD. Published May 14. Available: http://www.oecd-ilibrary.org/social-issues-migration-health/the-risk-of-automation-for-jobs-in-oecd-countries_5jlz9h56dvq7-en. [Accessed May 20 2016].

ARTICLE29. 2010. *Opinion 3/2010 on the principle of accountability*. European Commission. Published July. Available: http://ec.europa.eu/justice/policies/privacy/docs/wpdocs/2010/wp173_en.pdf. [Accessed June 6 2015].

ASKITAS, N. & ZIMMERMANN, K. F. 2015. *The Internet as a Data Source for Advancement in Social Sciences*. Forschungsinstitut zur Zukunft der Arbeit. Published February. Available: http://ftp.iza.org/dp8899.pdf. [Accessed March 4 2016].

BARANIUK, C. 2016. *DDoS: Website-crippling cyber-attacks to rise in 2016*. British Broadcasting Corporation (BBC). Published January 27. Available: http://www.bbc.co.uk/news/technology-35376327. [Accessed January 27 2016].

BAROCAS, S. & SELBST, A. D. 2016. *Big Data's Disparate Impact*. California Law Review, 104, 3, 671-732. Available: http://www.californialawreview.org/wp-content/uploads/2016/06/2Barocas-Selbst.pdf.

BARTLETT, J. & GRABBE, H. 2015. *e-democracy in the EU: the opportunities for digital politics to re-engage voters and the risks of disappointment*. Demos. Published December. Available: http://www.demos.co.uk/wp-content/uploads/2015/12/E-Democracy-in-the-EU.pdf. [Accessed August 24 2016].

BARTON, D., FARRELL, D. & MOURSHED, M. 2013. *Education to employment: Designing a system that works*. McKinsey & Company. Published January. Available: http://www.mckinsey.com/industries/social-sector/our-insights/education-to-employment-designing-a-system-that-works. [Accessed February 2 2014].

BBC. 2003. *Spy cameras in every class*. British Broadcasting Corporation (BBC). Available: http://news.bbc.co.uk/1/hi/education/3095885.stm. [Accessed July 27 2003].

BBC. 2004a. *Camera phones threat to privacy*. British Broadcasting Corporation (BBC). Available: http://news.bbc.co.uk/1/hi/technology/4017225.stm. [Accessed November 17 2004].

BBC. 2004b. *Pupils face camera phone ban*. British Broadcasting Corporation (BBC). Available: http://news.bbc.co.uk/1/hi/scotland/3524913.stm. [Accessed March 3 2004].

BBC. 2016. *PFI schools built in Scotland 'owned by offshore firms'*. British Broadcasting Corporation (BBC). Published August 22. Available: http://www.bbc.co.uk/news/uk-scotland-37135611. [Accessed August 22 2016].

BBW. 2016. *Classroom Management Software – Another Brick in the Wall? How schools use software to monitor pupils* Big Brother Watch (UK). Published November. Available: https://www.bigbrotherwatch.org.uk/wp-content/uploads/2016/11/Classroom-Management-Software-Another-Brick-in-the-Wall.pdf. [Accessed November 8 2016].

BECKMANN, C. 2003. *Pisa-Test in Mathematik: Teilnahme ist diesmal Pflicht*. Berliner Morgenpost. Published April 24. Available: http://www.morgenpost.de/printarchiv/berlin/article102166959/Pisa-Test-in-Mathematik-Teilnahme-ist-diesmal-Pflicht.html. [Accessed March 19 2017].

BERENDT, B., DE PAOLI, S., LAING, C., et al. 2014. *Roadmap for NIS education programmes in Europe*. ENISA - European Union Agency for Network and Information Security. Published October. Available: https://www.enisa.europa.eu/publications/roadmap-for-nis-education-programmes-in-europe. [Accessed November 5 2015].

BERENDT, B. & PREIBUSCH, S. 2014. *Better decision support through exploratory discrimination-aware data mining: foundations and empirical evidence*. Artificial Intelligence and Law*,* 22*,* 2*,* 175-209. Available: DOI: 10.1007/s10506-013-9152-0.

BIGGE, R. 2006. *The cost of (anti-)social networks: Identity, agency and neo-luddites*. First Monday*,* 11*,* 12*,* online. Available: http://journals.uic.edu/ojs/index.php/fm/article/view/1421.

BIRD, J. 2016. *More pupils are told to 'bring your own device' as school budget cuts bite*. Financial Times (London). Published May 4. Available: http://www.ft.com/cms/s/2/c51f9ee0-f744-11e5-96db-fc683b5e52db.html. [Accessed May 4 2016].

BJÖRKLUND, F. 2016. *E-government and moral citizenship: the case of Estonia*. Citizenship Studies*,* 20*,* 6-7*,* 914-931. Available: http://dx.doi.org/10.1080/13621025.2016.1213222.

BLOOM, B. S. 1984. *The 2 Sigma Problem: The Search for Methods of Group Instruction as Effective as One-to-One Tutoring* Educational Researcher*,* 13*,* 6*,* 4-16. Available: https://www.jstor.org/stable/1175554?seq=1#page_scan_tab_contents.

BOALER, J. 1997. *Setting, social class and survival of the quickest*. British Educational Research Journal*,* 23*,* 5*,* 575-595. Available: http://onlinelibrary.wiley.com/doi/10.1080/0141192970230503/abstract.

BOALER, J., WILIAM, D. & BROWN, M. 2000. *Students' experiences of ability grouping-disaffection, polarisation and the construction of failure*. British Educational Research Journal*,* 26*,* 5*,* 631-648. Available: http://www.informaworld.com/10.1080/713651583.

BOESKENS, L. 2016. *Regulating Publicly Funded Private Schools: A Literature Review on Equity and Effectiveness* OECD. Published November 11. Available: http://www.oecd-ilibrary.org/education/regulating-publicly-funded-private-schools_5jln6jcg80r4-en. [Accessed November 25 2016].

BOUCHER, P. 2017. *In-Depth Analysis : How blockchain technology could change our lives*. European Parliamentary Research Service. Published February. Available: http://www.europarl.europa.eu/RegData/etudes/IDAN/2017/581948/EPRS_IDA(2017)581948_EN.pdf. [Accessed March 26 2017].

BOYD, D. 2008. *Taken Out of Context: American Teen Sociality in Networked Publics*. University of Berkeley, California. Available: http://www.danah.org/papers/TakenOutOfContext.pdf [Accessed March 3 2017].

BOYD, D. & CRAWFORD, K. 2011. *Six Provocations for Big Data*. SSRN. Published September 21. Available: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1926431. [Accessed July 12 2013].

BOYD, D. & CRAWFORD, K. 2012. *Critical Questions for Big Data*. Information, Communication & Society, 15, 5, 662-679. Available: DOI: 10.1080/1369118X.2012.678878.

BRETT, C. 2015. *My life under Estonia's digital government*. The Register. Published June 2. Available: https://www.theregister.co.uk/2015/06/02/estonia/. [Accessed January 4 2017].

BRITTON, J., DEARDEN, L., SHEPHARD, N., et al. 2016. *How English domiciled graduate earnings vary with gender, institution attended, subject and socio-economic background*. Institute for Fiscal Studies. Published April 13. Available: http://www.ifs.org.uk/publications/8233. [Accessed April 20 2016].

BRITTON, J., SHEPHARD, N. & VIGNOLES, A. 2015. *Comparing sample survey measures of English earnings of graduates with administrative data during the Great Recession*. Institute for Fiscal Studies. Published September 23. Available: http://www.ifs.org.uk/uploads/publications/wps/WP201528.pdf. [Accessed July 5 2016].

BROOMFIELD, M. 2016. *Teenage cyberbullying victim mocked on Facebook days after suicide in front of parents*. Independent (London). Published December 14. Available: http://www.independent.co.uk/news/world/americas/facebook-cyberbullying-victim-rebecca-ann-sedwick-after-suicide-fake-profiles-florida-a7473606.html. [Accessed December 19 2016].

BUDGINAITĖ, I., SIAROVA, H., STERNADEL, D., et al. 2016. *Policies and practices for equality and inclusion in and through education*. European Commission. Available: https://bookshop.europa.eu/en/policies-and-practices-for-equality-and-inclusion-in-and-through-education-pbNC0415453/. [Accessed February 1 2017].

BUNDESVERFASSUNGSGERICHT. 1983. *BVerfGE 65, 1—Volkszählung. Urteil des Ersten Senats vom 15. Dezember 1983 auf die mündliche Verhandlung vom 18. und 19. Oktober 1983-1 BvR 209, 269, 362, 420, 440, 484/83 in den Verfahren über die Verfassungsbeschwerden*. Telemedicus.info. Available: https://www.telemedicus.info/urteile/Datenschutzrecht/88-BVerfG-Az-1-BvR-209,-269,-362,-420,-440,-48483-Volkszaehlungsurteil.html. [Accessed March 19 2018].

BURNS, B. 2016a. *Big Data's Coming Of Age In Higher Education*. Forbes.com. Published January 29. Available: http://www.forbes.com/sites/schoolboard/2016/01/29/big-datas-coming-of-age-in-higher-education/. [Accessed February 8 2016].

BURNS, J. 2016b. *Online criticism from parents wearing down teachers*. British Broadcasting Corporation (BBC). Published November 11. Available: http://www.bbc.co.uk/news/education-37940465. [Accessed November 11 2016].

BURNS, J. 2016c. *Social media harms moral development, parents say*. British Broadcasting Corporation (BBC). Published July 18. Available: http://www.bbc.com/news/education-36824176. [Accessed July 18 2016].

BURNS, T. & KÖSTER, F. (eds.) 2016. *Governing Education in a Complex World.* Paris: OECD.

BUTTARELLI, G. 2016. *Keynote address by Giovanni Buttarelli to the Delphi Economic Forum: Living in a future Big Data world: can prosperity, freedom and fundamental rights be reconciled?* European Data Protection Supervisor. Published February 27. Available: https://secure.edps.europa.eu/EDPSWEB/webdav/site/mySite/shared/Documents/EDPS/Publications/Speeches/2016/16-02-26_Delphi_Speech_EN.pdf. [Accessed February 29 2016].

CALUDE, C. S. & LONGO, G. 2016. *The Deluge of Spurious Correlations in Big Data*. Foundations of Science*,* 1-18. Available: http://link.springer.com/article/10.1007%2Fs10699-016-9489-4.

CANTINI, C., CHELLINI, C. & SAGRI, M.-T. 2016. *Big Data Analytics National Educational System Monitoring and Decision Making*. World Journal of Social Science Research*,* 3*,* 2. Available: http://www.scholink.org/ojs/index.php/wjssr/article/view/568.

CAPILUPPI, A. & BARAVALLE, A. 2010. *Matching Demand and Offer in On-line Provision: a Longitudinal Study of Monster.com*. WSE 2010 Proceedings the 12th IEEE International Symposium on Web Systems Evolution. Published September. Available: http://roar.uel.ac.uk/995/1/Capiluppi,%20A%20(2010)%20WSE.pdf. [Accessed December 3 2016].

CHAMPAIGN, J., COLVIN, K. F., LIU, A., et al. 2014. *Correlating Skill and Improvement in 2 MOOCs with a Student's Time on Tasks* L@S 2014 • Student Skills and Behavior. Published March. Available: https://edocs.uis.edu/jcham4/www/p11-champaign.pdf. [Accessed March 5 2016].

CLAROS, E. & DAVIES, R. 2016. *Economic impact of Big Data*. European Parliamentary Research Service. Published September 29. Available: https://epthinktank.eu/2016/09/29/economic-impact-of-big-data/. [Accessed October 11 2016].

CMEC. 2010. *A Framework for Statistics on Learning and Education in Canada*. Council of Ministers of Education, Canada. Available: http://www.cmec.ca/publications/lists/publications/attachments/257/cesc-data-framework-sept2010.pdf. [Accessed October 4 2016].

COLVIN, K. F., CHAMPAIGN, J., LIU, A., et al. 2014. *Learning in an Introductory Physics MOOC: All Cohorts Learn Equally, Including an On-Campus Class*. International Review of Research in Open and Distance Learning*,* 15*,* 4*,* online. Available: http://www.irrodl.org/index.php/irrodl/article/view/1902/3009.

COMMISSION. 1993. *Growth, competitiveness and employment: the challenges and ways forward into the 21st century: White Paper COM(93) 700 final.* Brussels: European Commission. Available: http://aei.pitt.edu/1139/.

COMMISSION. 1994. *Europe's way to the Information Society: an action plan: COM(94) 347 Final.* Brussels: European Commission. Available: http://aei.pitt.edu/947/.

COMMISSION. 2006a. *Efficiency and equity in European education and training systems {SEC(2006) 1096} /* COM/2006/0481 final */*. European Commission. Published September 8. Available: http://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX%3A52006DC0481. [Accessed February 3 2007].

COMMISSION. 2006b. *History of European co-operation in education*. European Commission. Available: http://bookshop.europa.eu/en/the-history-of-european-cooperation-in-education-and-training-pbNC6505311/. [Accessed May 11 2011].

COMMISSION. 2007. *Communication from the Commission to the European Parliament and the Council on Promoting Data Protection by Privacy Enhancing Technologies (PETs). /* COM/2007/0228 final */*. European Commission. Published May 2. Available: http://eur-lex.europa.eu/legal-content/en/TXT/?uri=CELEX%3A52007DC0228. [Accessed January 4 2008].

COMMISSION. 2009. *Education and Training 2020 (ET 2020)*. European Commission. Published October 23. Available: http://europa.eu/legislation_summaries/education_training_youth/general_framework/ef0016_en.htm. [Accessed November 26 2013].

COMMISSION. 2010. *Towards interoperability for European public services*. European Commission. Published December 16. Available: http://europa.eu/rapid/pressReleasesAction.do?reference=IP/10/1734&format=HTML&aged=0&language=EN&guiLanguage=en. [Accessed December 16 2010].

COMMISSION. 2012a. *Commission proposes a comprehensive reform of data protection rules to increase users' control of their data and to cut costs for businesses*. European Commission. Published January 25. Available: http://europa.eu/rapid/pressReleasesAction.do?reference=IP/12/46&format=HTML&aged=0&language=EN&guiLanguage=en. [Accessed January 25 2012].

COMMISSION. 2012b. *Rethinking Education: Investing in skills for better socio-economic outcomes*. European Commission. Published November 20. Available: http://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1389776578033&uri=CELEX:52012DC0669. [Accessed November 21 2012].

COMMISSION. 2013a. *Ethics for researchers* European Commission. Available: http://doi.org/10.2777/7491. [Accessed January 5 2014].

COMMISSION. 2013b. *Investing in our future: Good examples of EU funded projects*. European Commission. Published February 7. Available: http://europa.eu/rapid/press-release_MEMO-13-74_en.htm?locale=en. [Accessed February 7 2013].

COMMISSION. 2013c. *Opening up Education: Innovative teaching and learning for all through new Technologies and Open Educational Resources*. European Commission. Published September 25. Available: http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=COM:2013:0654:FIN:EN:PDF. [Accessed September 26 2013].

COMMISSION. 2014a. *Commission urges governments to embrace potential of Big Data*. European Commission. Published July 2. Available: http://europa.eu/rapid/press-release_IP-14-769_en.htm?locale=en. [Accessed July 2 2014].

COMMISSION. 2014b. *EU high level group calls for targeted funding to boost use of new technologies in higher education*. European Commission. Published October 22. Available: http://europa.eu/rapid/press-release_IP-14-1188_en.htm?locale=en. [Accessed October 22 2014].

COMMISSION. 2014c. *Factsheet on the "Right to be Forgotten" ruling (C-131/12)*. European Commission. Available: http://ec.europa.eu/justice/data-protection/files/factsheets/factsheet_data_protection_en.pdf. [Accessed January 8 2015].

COMMISSION. 2014d. *Towards a thriving data-driven economy (COM/2014/0442 final)*. European Commission. Published July 2. Available: http://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1404888011738&uri=CELEX:52014DC0442. [Accessed August 6 2014].

COMMISSION. 2015. *Draft Joint Report on the implementation of the strategic framework for European cooperation in education and training (ET 2020) – Communication from the Commission to the Council*. European Commission. Published February 3. Available: http://ec.europa.eu/smart-regulation/roadmaps/docs/2014_eac_008_joint_report_education_and_training.en.pdf. [Accessed May 6 2015].

COMMISSION. 2016a. *Education & training in Europe: investment on the rebound but inclusion remains a challenge*. European Commission. Published November 7. Available: http://europa.eu/rapid/press-release_IP-16-3577_en.htm?locale=en. [Accessed November 8 2016].

COMMISSION. 2016b. *Helping educational organisations to go digital*. European Commission. Published April 28. Available: https://ec.europa.eu/jrc/en/news/helping-educational-organisations-go-digital?r=dnl. [Accessed May 4 2016].

COMMISSION. 2016c. *Proposal for a Council Recommendation on establishing a Skills Guarantee*. European Commission. Published June 10. Available: http://ec.europa.eu/social/BlobServlet?docId=15621&langId=en. [Accessed June 10 2016].

COMMISSION. 2016d. *Ten actions to help equip people in Europe with better skills*. European Commission. Published June 10. Available: http://europa.eu/rapid/press-release_IP-16-2039_en.htm?locale=en. [Accessed June 10 2016].

COMMISSION. 2017a. *Big Data Value Public-Private Partnership*. European Commission. Published January 24. Available: https://ec.europa.eu/digital-single-market/en/big-data-value-public-private-partnership. [Accessed February 28 2017].

COMMISSION. 2017b. *European Interoperability Framework: Commission presents new guidance for digital public services*. European Commission. Published March 23. Available: http://europa.eu/rapid/press-release_IP-17-702_en.htm?locale=en. [Accessed March 23 2017].

COMMISSION. 2017c. *Taking data protection into a digital and globalised era: Joint Statement by Vice-President Ansip and Commissioner Jourová ahead of the 2017 Data Protection day*. European Commission. Published January 27. Available: http://europa.eu/rapid/press-release_STATEMENT-17-154_en.htm?locale=en. [Accessed January 28 2017].

COMMISSION. 2017d. *What can big data do for you ?* European Commission. Published March 1. Available: https://ec.europa.eu/digital-single-market/en/what-big-data-can-do-you. [Accessed March 8 2017].

COMMONS. 2014. *Responsible Use of Data*. House of Commons, Science and Technology Select Committee. Published November 28. Available: http://www.publications.parliament.uk/pa/cm201415/cmselect/cmsctech/245/24502.htm. [Accessed December 5 2014].

COMMONS. 2016. *Training new teachers* House of Commons, Public Accounts Select Committee. Published June 10. Available:

http://www.publications.parliament.uk/pa/cm201617/cmselect/cmpubacc/73/7302.htm. [Accessed June 10 2016].

CONDLIFFE, J. 2016. *DeepMind Will Use AI to Streamline Targeted Cancer Treatment*. Technology Review. Published August 31. Available: https://www.technologyreview.com/s/602277/deepmind-will-use-ai-to-streamline-targeted-cancer-treatment/?set=602270. [Accessed August 31 2016].

COOPER, P. 2016. *Department of Justice Wages War On Free Education*. Forbes.com. Published September 21. Available: http://www.forbes.com/sites/prestoncooper2/2016/09/21/department-of-justice-wages-war-on-free-education/#6863c351331c. [Accessed October 17 2016].

COR. 2014. *CoR calls for a pan-European strategy in the recognition and validation of informal and non-formal learning*. Committee of the Regions of the European Union Published December 4. Available: http://cor.europa.eu/en/news/Pages/informal-and-non-formal-learning.aspx. [Accessed December 4 2014].

CORMEN, T. H., LEISERSON, C. E., RIVEST, R. L., et al. 2009. *Introduction to Algorithms, Third Edition.* Cambridge, MA: MIT Press.

COUNCIL. 2011. *Council conclusions on early childhood education and care: providing all our children with the best start for the world of tomorrow*. Council of The European Union. Published May 20. Available: https://www.consilium.europa.eu/uedocs/cms_data/docs/pressdata/en/educ/122123.pdf. [Accessed February 4 2013].

COUNCIL. 2015a. *Council conclusions on the role of early childhood education and primary education in fostering creativity, innovation and digital competence*. Council of The European Union. Published May 27. Available: http://publications.europa.eu/en/publication-detail/-/publication/dab551ab-0432-11e5-a4c8-01aa75ed71a1/language-en/format-PDFA1A/source-1730183. [Accessed June 5 2015].

COUNCIL. 2015b. *Draft 2015 Joint Report of the Council and the Commission on the implementation of the Strategic framework for European cooperation in education and training (ET2020): New priorities for European cooperation in education and training. {SWD(2015) 161 final}*. Council of The European Union. Published August 26. Available: http://ec.europa.eu/education/documents/et-2020-draft-joint-report-408-2015_en.pdf. [Accessed September 2 2015].

COUNCIL. 2016a. *Council conclusions on the role of the youth sector in an integrated and cross-sectoral approach to preventing and combating violent radicalisation of young people*. Council of The European Union. Published May 30. Available: http://www.consilium.europa.eu/en/press/press-releases/2016/05/30-preventing-radicalisation/. [Accessed May 31 2016].

COUNCIL. 2016b. *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance)*. Council of The European Union. Published April 27. Available: http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=uriserv:OJ.L_.2016.119.01.0001.01.ENG&toc=OJ:L:2016:119:TOC. [Accessed May 30 2016].

CRAW, J. 2017. *Statistic of the Month: High Performance, Equity and Efficiency in Education*. National Center on Education and the Economy (USA). Published March

3. Available: http://ncee.org/2017/03/statistic-of-the-month-high-performance-equity-and-efficiency-in-education/. [Accessed March 10 2017].

D' ACQUISTO, G., DOMINGO-FERRER, J., KIKIRA, P., et al. 2015. *Privacy by design in big data* ENISA - European Union Agency for Network and Information Security. Published December. Available: https://www.enisa.europa.eu/publications/big-data-protection. [Accessed December 14 2016].

DALLA POZZA, V., DI PIETRO, A., MOREL, S., et al. 2016. *Cyberbullying among young people*. European Parliament. Published July. Available: http://www.europarl.europa.eu/RegData/etudes/STUD/2016/571367/IPOL_STU(2016)571367_EN.pdf. [Accessed October 2 2016].

DANEZIS, G., DOMINGO-FERRER, J., HANSEN, M., et al. 2015. *Privacy and Data Protection by Design*. ENISA - European Union Agency for Network and Information Security. Published January 12. Available: https://www.enisa.europa.eu/activities/identity-and-trust/library/deliverables/privacy-and-data-protection-by-design. [Accessed March 2 2017].

DARIES, J. P., REICH, J., WALDO, J., et al. 2014. *Privacy, Anonymity, and Big Data in the Social Sciences*. Communications of the ACM*,* 57*,* 9*,* 56-63. Available: http://cacm.acm.org/magazines/2014/9/177926-privacy-anonymity-and-big-data-in-the-social-sciences/fulltext.

DAVIET, B. 2016. *Revisiting the principle of education as a public good*. UNESCO. Published July. Available: http://www.unesco.org/ulis/cgi-bin/ulis.pl?catno=245306&set=0057ADC385_1_23&database=new2&gp=0&mode=e&ll=1. [Accessed August 12 2016].

DCMS. 2017. *UK Digital Strategy*. Department for Culture Media and Sport (UK). Published March 1. Available: https://www.gov.uk/government/publications/uk-digital-strategy. [Accessed March 1 2017].

DE MAURO, A., GRECO, M. & GRIMALDI, M. 2015. *What is big data? A consensual definition and a review of key research topics*. AIP Conference Proceedings*,* 1644*,* 1*,* 97-104. Available: http://doi.org/10.1063/1.4907823.

DE WOLF, R., VANDERHOVEN, E., BERENDT, B., et al. 2016. *Self-reflection on privacy research in social networking sites*. Behaviour & Information Technology*,* Online. Available: http://dx.doi.org/10.1080/0144929X.2016.1242653

DEDE, C. 2015. *Data-Intensive Research in Education: Current Work and Next Steps*. Computing Research Association Published November. Available: http://cra.org/wp-content/uploads/2015/10/CRAEducationReport2015.pdf. [Accessed March 14 2016].

DEDE, C., HO, A. & MITROS, P. 2016. *Big Data Analysis in Higher Education: Promises and Pitfalls*. Educause Review. Published August 22. Available: http://er.educause.edu/articles/2016/8/big-data-analysis-in-higher-education-promises-and-pitfalls. [Accessed November 21 2016].

DFE. 2016. *Graduate outcomes: longitudinal education outcomes (LEO) data*. Department for Education (UK). Published August 4. Available: https://www.gov.uk/government/statistics/graduate-outcomes-longitudinal-education-outcomes-leo-data. [Accessed August 16 2016].

DFE. 2017a. *Cloud software services: how schools should protect data*. Department for Education (UK). Published January 4. Available: https://www.gov.uk/government/publications/cloud-software-services-and-the-data-protection-act. [Accessed January 5 2017].

DFE. 2017b. *School performance and house prices: analysis*. Department for Education (UK). Published March 19. Available: https://www.gov.uk/government/publications/school-performance-and-house-prices-analysis. [Accessed March 20 2017].

DOBBIE, W. S. & FRYER JR, R. G. 2016. *Charter Schools and Labor Market Outcomes*. National Bureau of Economic Research (UDA). Available: http://papers.nber.org/papers/w22502. [Accessed December 12 2016].

DURSTON, S. 2014. *Intersectorality: The key to quality and equity in education*. Development Progress. Published July 18. Available: http://www.developmentprogress.org/blog/2014/07/18/intersectorality-key-quality-and-equity-education. [Accessed December 5 2016].

DUSI, S., MERCORIO, F. & MEZZANZANICA, M. 2015. *Big Data meets Job Vacancy: Trends, Challenges and Development Directions*. pp.31-45 *In:* LARSEN, C., RAND, S., SCHMID, A., et al. (eds.) *Big Data and the Complexity of Labour Market Policies: New Approaches in Regional and Local Labour Market Monitoring for Reducing Skills Mismatches.* Munich: Rainer Hampp Verlag.

DYSON, L., GOLDRICK, S., JONES, L., et al. 2010. *Equity in Education: Creating a fairer education system*. University of Manchester. Published April. Available: http://hummedia.manchester.ac.uk/institutes/cee/equity_in_education.pdf. [Accessed October 5 2016].

ECONOMIST. 2014. *Eyes on the classroom*. Economist. Published July 26. Available: http://www.economist.com/news/americas/21608640-close-education-gap-latin-america-must-produce-better-teachers-eyes-classroom. [Accessed August 4 2014].

ECONOMIST. 2016a. *Culture or policy? What the world can learn from the latest PISA test results*. Economist. Published December 10. Available: http://www.economist.com/news/international/21711247-reforming-education-slow-and-hard-eminently-possible-what-world-can-learn. [Accessed December 8 2016].

ECONOMIST. 2016b. *Data in schools: Heads in the clouds*. Economist. Published May 21. Available: http://www.economist.com/news/britain/21699187-pupil-data-are-too-hard-access-meaning-schools-are-losing-out-heads-clouds. [Accessed May 19 2016].

ECONOMIST. 2016c. *Education protest in Mexico: A battle to feed young minds*. Economist. Published July 9. Available: http://www.economist.com/news/americas/21701789-why-teachers-are-so-disobedient-battle-feed-young-minds. [Accessed July 7 2016].

ECONOMIST. 2016d. *Education reform: Teaching the teachers*. Economist. Published June 11. Available: http://www.economist.com/news/briefing/21700385-great-teaching-has-long-been-seen-innate-skill-reformers-are-showing-best. [Accessed June 9 2016].

ECONOMIST. 2016e. *Machine learning: Of prediction and policy*. Economist. Published August 20. Available: http://www.economist.com/news/finance-and-economics/21705329-governments-have-much-gain-applying-algorithms-public-policy. [Accessed August 18 2016].

ECONOMIST. 2016f. *A running start: Poor children fall behind early in life. Better pre-school education could help*. Economist. Published June 25. Available: http://www.economist.com/news/finance-and-economics/21701131-poor-

children-fall-behind-early-life-better-pre-school-education-could-help. [Accessed August 1 2016].

ECONOMIST. 2016g. *Teacher workload: All work and low pay*. Economist. Published April 16. Available: http://www.economist.com/news/britain/21696966-teachers-england-gripe-about-long-hoursand-justly-so-all-work-and-low-pay. [Accessed April 15 2016].

ECONOMIST. 2017. *Health care: Will artificial intelligence help to crack biology?* Economist. Published January 7. Available: http://www.economist.com/news/science-and-technology/21713828-silicon-valley-has-squidgy-worlds-biology-and-disease-its-sights-will. [Accessed January 8 2017].

EDPS. 2016. *EDPS-BEUC Conference on Big Data: individual rights and smart enforcement - 29 September 2016*. European Data Protection Supervisor. Published July 14. Available: https://secure.edps.europa.eu/EDPSWEB/edps/site/mySite/Events. [Accessed August 17 2016].

EDPS. 2017. *Big data and the Digital Clearing House*. European Data Protection Supervisor. Published January. Available: https://secure.edps.europa.eu/EDPSWEB/edps/Consultation/big_data. [Accessed January 27 2017].

EDTECHXGLOBAL. 2016. *Global Report Predicts EdTech Spend to Reach $252bn by 2020*. MarketWatch. Published May 25. Available: http://www.marketwatch.com/story/global-report-predicts-edtech-spend-to-reach-252bn-by-2020-2016-05-25-4203228. [Accessed June 5 2016].

EIB. 2016. *EU bank increases its presence in Hungary and supports upgrading of Hungary's educational infrastructure*. European Investment Bank. Published June 13. Available: http://www.eib.org/infocentre/press/releases/all/2016/2016-145-eu-bank-increases-its-presence-in-hungary-and-supports-upgrading-of-hungarys-educational-infrastructure.htm?media=rss&language=en. [Accessed June 13 2016].

EP. 2007. *Efficiency and equity in European education and training systems*. European Parliament. Published September 27. Available: http://www.europarl.europa.eu/sides/getDoc.do?Type=TA&Reference=P6-TA-2007-0417&language=EN. [Accessed January 5 2012].

EP. 2017. *European Parliament resolution of 14 March 2017 on fundamental rights implications of big data: privacy, data protection, non-discrimination, security and law-enforcement (2016/2225(INI))*. European Parliament. Published March 15. Available: http://www.europarl.europa.eu/sides/getDoc.do?pubRef=-//EP//TEXT+TA+P8-TA-2017-0076+0+DOC+XML+V0//EN&language=EN. [Accessed March 16 2017].

ESMA. 2016. *European Supervisory Authorities consult on Big Data*. Joint Committee of the European Supervisory Authorities. Published December 19. Available: https://www.esma.europa.eu/press-news/esma-news/european-supervisory-authorities-consult-big-data. [Accessed December 19 2016].

ESTAT. 2013. *Scheveningen Memorandum on "Big Data and Official Statistics"*. Eurostat. Published October 9. Available: https://ec.europa.eu/eurostat/cros/news/scheveningen-memorandum-big-data-and-official-statistics-adopted-essc_en. [Accessed December 12 2016].

ESTAT. 2014. *ESS Big Data Action Plan and Roadmap 1.0*. Eurostat. Published September 26. Available: https://ec.europa.eu/eurostat/cros/content/ess-big-data-action-plan-and-roadmap-10_en. [Accessed December 4 2015].

ESTAT. 2015. *Big Data: The Eurostat Perspective*. Eurostat. Published September. Available: https://ec.europa.eu/jrc/sites/jrcsh/files/j3-bigdata-02-Wirthmann.pdf. [Accessed March 4 2016].

ESTAT. 2016a. *1 out of 4 internet users in the EU experienced security related problems in 2015*. Eurostat. Published February 8. Available: http://ec.europa.eu/eurostat/web/products-press-releases/-/4-08022016-AP. [Accessed February 8 2016].

ESTAT. 2016b. *ESSnet Big Data*. Eurostat. Published November 10. Available: https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/index.php/ESSnet_Big_Data. [Accessed January 6 2017].

ESTAT. 2016c. *Statistical approaches to the measurement of skills* Eurostat. Available: https://bookshop.europa.eu/en/statistical-approaches-to-the-measurement-of-skills-pbKSTC16023/?CatalogCategoryID=iEKep2Ix3hEAAAEud3kBgSLq. [Accessed January 18 2017].

ESTAT. 2016d. *Women teachers largely over-represented in primary education in the EU*. Eurostat. Published October 4. Available: http://ec.europa.eu/eurostat/documents/2995521/7672738/3-04102016-BP-EN.pdf/34f53c26-9e44-4983-bfc9-b98019af6cd7. [Accessed October 4 2016].

EUN. 2013. *Survey of Schools: ICT in Education. Executive Summary*. European Schoolnet. Published February. Available: http://www.eun.org/c/document_library/get_file?uuid=9be81a75-c868-4558-a777-862ecc8162a4&groupId=43887. [Accessed January 4 2016].

EUROFOUND. 2016. *Foundation Seminar Series (FSS) 2016 - The impact of digitalisation on work: Building up national agendas for better implementation of digital changes - Session 1*. European Foundation for the Improvement of Living and Working Conditions. Published May 26. Available: http://www.eurofound.europa.eu/events/working-conditions-labour-market-industrial-relations-law-and-regulation-quality-of-life-social/foundation-seminar-series-fss-2016-the-impact-of-digitalisation-on-work-building-up-national-agendas-0. [Accessed May 26 2016].

EURYDICE. 2015a. *Structural indicators for monitoring education and training systems in Europe - 2015* European Commission. Published November 12. Available: http://eacea.ec.europa.eu/education/eurydice/documents/thematic_reports/190EN.pdf. [Accessed December 15 2015].

EURYDICE. 2015b. *The Teaching Profession in Europe: Practices, Perceptions, and Policies*. European Commission. Published June 25. Available: https://webgate.ec.europa.eu/fpfis/mwikis/eurydice/index.php/Publications:The_Teaching_Profession_in_Europe:_Practices,_Perceptions,_and_Policies. [Accessed July 2 2015].

EURYDICE. 2016a. *Compulsory education in Europe 2016/17*. Education, Audiovisual & Culture Executive Agency. Available: https://bookshop.europa.eu/en/compulsory-education-in-europe-2016-17-pbECAP16001/?CatalogCategoryID=iEKep2Ix3hEAAAEud3kBgSLq. [Accessed November 25 2016].

EURYDICE. 2016b. *The Organisation of School Time in Europe. Primary and General Secondary Education – 2016/17*. European Commission. Published September 6.

Available: https://webgate.ec.europa.eu/fpfis/mwikis/eurydice/index.php/Publications:The_Organisation_of_School_Time_in_Europe._Primary_and_General_Secondary_Education_%E2%80%93_2016/17. [Accessed September 26 2016].

EURYDICE. 2016c. *The Organisation of the Academic Year in Europe – 2016/17*. European Commission. Published September 6. Available: https://webgate.ec.europa.eu/fpfis/mwikis/eurydice/index.php/Publications:The_Organisation_of_the_Academic_Year_in_Europe_%E2%80%93_2016/17. [Accessed September 26 2016].

EURYDICE. 2016d. *Recommended Annual Instruction Time in Full-time Compulsory Education in Europe 2015/16*. European Commission. Published June 9. Available: https://webgate.ec.europa.eu/fpfis/mwikis/eurydice/index.php/Publications:Recommended_Annual_Instruction_Time_in_Full-time_Compulsory_Education_in_Europe_2015/16. [Accessed July 1 2016].

EURYDICE. 2016e. *Structural indicators for monitoring education and training systems in Europe 2016* European Commission. Published October. Available: http://bookshop.europa.eu/is-bin/INTERSHOP.enfinity/WFS/EU-Bookshop-Site/en_GB/-/EUR/ViewPublication-Start?PublicationKey=EC0116905. [Accessed January 6 2017].

EURYDICE. 2017a. *Structural Indicators for Monitoring Education and Training Systems in Europe 2016 – Thematic Overviews*. European Commission. Published February 16. Available: https://webgate.ec.europa.eu/fpfis/mwikis/eurydice/index.php/Publications:Structural_Indicators_for_Monitoring_Education_and_Training_Systems_in_Europe_2016_%E2%80%93_Thematic_Overviews. [Accessed February 16 2017].

EURYDICE. 2017b. *Support Mechanisms for Evidence-based Policy-Making in Education* Education, Audiovisual & Culture Executive Agency. Published January 24. Available: https://webgate.ec.europa.eu/fpfis/mwikis/eurydice/index.php/Publications:Support_Mechanisms_for_Evidence-based_Policy-Making_in_Education. [Accessed January 27 2017].

EVARISTO, T. 2014. *Portuguese Information Systems for Education Monitoring*. OECD. Published July. Available: https://www.oecd.org/edu/ceri/Teresa%20Evaristo_Portuguese%20Information%20Systems%20for%20Education%20Monitoring.pdf. [Accessed June 3 2016].

FASSBENDER, I. & GIAMBRONE, A. 2015. *When Big Data Meets the Blackboard*. The Atlantic. Published June 22. Available: http://www.theatlantic.com/education/archive/2015/06/big-data-student-privacy/396452/. [Accessed November 21 2016].

FERGUSON, R., BRASHER, A., CLOW, D., et al. 2016. *Research Evidence on the Use of Learning Analytics: Implications for Education Policy*. Joint Research Centre (European Commission). Available: https://ec.europa.eu/jrc/en/publication/eur-scientific-and-technical-research-reports/research-evidence-use-learning-analytics-implications-education-policy. [Accessed December 2 2016].

FERREIRA, J. 2015. *Big Data in the Education System*. Brainscape.com. Published September 8. Available: https://www.brainscape.com/blog/2014/11/big-data-education-system/. [Accessed February 3 2016].

FIEDLER, E. D., LANGE, R. E. & WINEBRENNER, S. 2002. *In search of reality: Unraveling the myths about tracking, ability grouping, and the gifted*. Roeper Review, 24, 3, 108-111. Available: http://dx.doi.org/10.1080/02783190209554142

FRICK, T. W. 2016. *The Theory of Totally Integrated Education: TIE*. Indiana University. Published October 12. Available: http://educology.indiana.edu/Frick/TIEtheory.pdf. [Accessed December 6 2016].

GAEBEL, M., HAUSCHILDT, K., MÜHLECK, K., et al. 2012. *Tracking Learners' and Graduates' Progression Paths TRACKIT*. European University Association. Available: http://www.eua.be/Libraries/publications-homepage-list/EUA_Trackit_web.pdf?sfvrsn=2. [Accessed June 4 2016].

GAO. 2014. *Child Welfare; Federal Agencies Can Better Support State Efforts to Prevent and Respond to Sexual Abuse by School Personnel*. US Government Accountability Office. Published January. Available: http://www.gao.gov/assets/670/660375.pdf. [Accessed June 4 2016].

GAO, B. & BERENDT, B. 2014. *Friends and circles – A design study for contact management in egocentric online social networks*. pp.1-34 *In:* KAWASH, J. (ed.) *Online Social Media Analysis and Visualization.* New York: Springer.

GARTNER. 2014. *Hype Cycle for Education, 2014*. Gartner Inc. Published July 23. Available: https://www.gartner.com/doc/2806424/hype-cycle-education-. [Accessed August 8 2014].

GARTNER. 2016. *Hype Cycle for Education, 2016*. Gartner Inc. Published July 1. Available: https://www.gartner.com/doc/3364119/hype-cycle-education-. [Accessed November 2 2016].

GEE, O. 2015. *The troubles with French universities laid bare*. The Local FR. Published October 9. Available: http://www.thelocal.fr/20151009/eight-troubles-with-french-universities-laid-bare. [Accessed October 19 2015].

GORE, A. 1994. *Speech delivered at the Information Superhighway Summit at UCLA January 11, 1994*. White House (USA).

GRAZIANO, C. 2003. *Learning to Live With Biometrics*. Wired.com. Published September 9. Available: https://citizenlab.org/2003/09/learning-to-live-with-biometrics/. [Accessed September 14 2003].

GROS, B. 2016. *The design of smart educational environments*. Smart Learning Environments, 3, 15, 1-11. Available: https://slejournal.springeropen.com/articles/10.1186/s40561-016-0039-x.

GSR. 2016. *Social Media Research Guidance; using social media for social research*. Government Social Research Profession (UK). Published May 24. Available: https://www.gov.uk/government/publications/social-media-research-guidance-using-social-media-for-social-research. [Accessed August 4 2016].

GUTWIRTH, S. & DE HERT, P. 2006. *Privacy, data protection and law enforcement. Opacity of the individual and transparency and power*. pp.61-104 *In:* CLAES, E., DUFF, A. & GUTWIRTH, S. (eds.) *Privacy and the Criminal Law.* Antwerp/Oxford: Intersentia.

HAGGERTY, K. D. & ERICSON, R. V. 2000. *The surveillant assemblage*. British Journal of Sociology, 51, 4, 605–622.

HAJIAN, S., BONCHI, F. & CASTILLO , C. 2016. *Algorithmic bias: from discrimination discovery to fairness-aware data mining*. Francescobonchi.com. Published March. Available: http://francescobonchi.com/algorithmic_bias_tutorial.html. [Accessed March 19 2017].

HAKE, R. 1998. *Interactive-Engagement Versus Traditional Methods: A Six-Thousand-Student Survey of Mechanics Test Data for Introductory Physics Courses*. American Journal of Physics, 66, 1, 64-74.

HALLINAN, M. T. 1994. *Tracking: From Theory to Practice*. Sociology of Education, 67, 2, 79-84. Available: http://www.jstor.org/stable/2112697.

HAMPSON, M., PATTON, A. & SHANKS, L. 2016. *10 ideas for 21st Century Education*. Innovation Unit. Available: http://www.innovationunit.org/resources/10-ideas-21st-century-education. [Accessed January 25 2017].

HANDEL, M. J. 2012. *Trends in Job Skill Demands in OECD Countries*. OECD. Published November 28. Available: http://dx.doi.org/10.1787/5k8zk8pcq6td-en. [Accessed January 6 2013].

HANUSHEK, E. A. & WÖßMANN, L. 2006. *Does educational tracking affect performance and inequality? Differences-in-differences evidence across countries*. The Economic Journal, 116, 510, C63-C76. Available: DOI: 10.1111/j.1468-0297.2006.01076.x.

HAR CARMEL, Y. 2016. *Regulating "big data education" in Europe: lessons learned from the US*. Internet Policy Review, 5, 1, online. Available: https://policyreview.info/articles/analysis/regulating-big-data-education-europe-lessons-learned-us.

HARDT, M., PRICE, E. & SREBRO, N. 2017. *Equality of Opportunity in Supervised Learning*. University of Chicago. Published October 7. Available: http://ttic.uchicago.edu/~nati/Publications/HardtPriceSrebro2016.pdf. [Accessed December 5 2016].

HARGREAVES, A. & SHIRLEY, D. 2009. *The Fourth Way: The Inspiring Future for Educational Change.* Thousand Oaks, CA: Corwin Press.

HARRIS, M. 2016. *What to Do When a Robot Is the Guilty Party*. Technology Review. Published May 25. Available: https://www.technologyreview.com/s/601563/what-to-do-when-a-robot-is-the-guilty-party/#/set/id/601579/. [Accessed May 26 2016].

HARVEN, M. 2013. *Big Data in Education: Key Players in Data Analytics Weigh In*. Edtechtimes.com. Published December 18. Available: https://edtechtimes.com/2013/12/18/big-data-perspective-big-players-data-analytics/. [Accessed January 3 2015].

HATTIE, J. 2015. *The applicability of Visible Learning to higher education*. Scholarship of Teaching and Learning in Psychology, 1, 1, 79-91. Available: http://dx.doi.org/10.1037/stl0000021.

HEROLD, B. 2016. *Digital Tools Evolving to Track Students' Emotions, Mindsets*. Education Week. Published January 11. Available: http://www.edweek.org/ew/articles/2016/01/13/digital-tools-evolving-to-track-students-emotions.html. [Accessed November 21 2016].

HIPPE, R., ARAÚJO, L. & DA COSTA, P. D. 2016. *Equity in education in Europe* Joint Research Centre (European Commission). Available: https://bookshop.europa.eu/en/equity-in-education-in-europe-pbLFNA28285/?CatalogCategoryID=iEKep2Ix3hEAAAEud3kBgSLq. [Accessed January 30 2017].

HML. 2015. *The Iceberg Effect: An International Look at Often-Overlooked Education Indicators*. The Horace Mann League of the U.S.A., and the National Superintendents Roundtable. Published January. Available: http://www.superintendentsforum.org/wp-content/uploads/2015/01/Released-Iceberg-Effect.pdf. [Accessed January 3 2017].

HOLLAND, D., LIADZE, I., RIENZO, C., et al. 2013. *The relationship between graduates and economic growth across countries*. Department for Business, Innovation and

Skills (UK). Published August. Available: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/229492/bis-13-858-relationship-between-graduates-and-economic-growth-across-countries.pdf. [Accessed June 4 2014].

ICEF. 2016. *MOOC enrolment surpassed 35 million in 2015*. ICEF Monitor. Published January 5. Available: http://monitor.icef.com/2016/01/mooc-enrolment-surpassed-35-million-in-2015/. [Accessed March 3 2016].

ICO. 2017. *Overview of the General Data Protection Regulation (GDPR)*. Information Commission (UK). Published March 3. Available: https://ico.org.uk/media/for-organisations/data-protection-reform/overview-of-the-gdpr-1-7.pdf. [Accessed March 19 2017].

INAMORATO DOS SANTOS, A., PUNIE, Y. & CASTAÑO MUÑOZ, J. 2016. *Opening up Education: A Support Framework for Higher Education Institutions*. Institute for Prospective Technological Studies. Available: https://ec.europa.eu/jrc/en/publication/eur-scientific-and-technical-research-reports/opening-education-support-framework-higher-education-institutions. [Accessed August 14 2016].

JONES, R., PYKETT, J. & WHITEHEAD, M. 2013. *Changing Behaviours. On the Rise of the Psychological State.* Cheltenham UK: Edward Elgar Publishing.

KAMENTEZ, A. 2014. *What Parents Need To Know About Big Data And Student Privacy*. Npr.org. Published April 28. Available: http://www.npr.org/sections/alltechconsidered/2014/04/28/305715935/what-parents-need-to-know-about-big-data-and-student-privacy. [Accessed December 3 2016].

KAMENTEZ, A. 2015. *Nonacademic Skills Are Key To Success. But What Should We Call Them?* nprED. Published May 28. Available: http://www.npr.org/sections/ed/2015/05/28/404684712/non-academic-skills-are-key-to-success-but-what-should-we-call-them. [Accessed December 12 2016].

KELION, L. 2015. *Schools given Dropbox guidance after Safe Harbour warning*. British Broadcasting Corporation (BBC). Published October 28. Available: http://www.bbc.co.uk/news/technology-34646146. [Accessed October 28 2015].

KIRBY, P. & CULLINANE, C. 2017. *Science Shortfall* Sutton Trust. Published January 27. Available: http://www.suttontrust.com/researcharchive/science-shortfall/. [Accessed February 9 2017].

KITCHIN, R. 2014. *The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences.* New York: Sage.

KNIGHT, W. 2015. *Inside Amazon's Warehouse, Human-Robot Symbiosis*. Technology Review. Published July 7. Available: http://www.technologyreview.com/news/538601/inside-amazons-warehouse-human-robot-symbiosis/. [Accessed July 7 2015].

KOEDINGER, K. R., BAKER, R. S. J. D., CUNNINGHAM, K., et al. 2010. *A Data Repository for the EDM Community: The PSLC DataShop*. pp.43-53 *In:* ROMERO, C., VENTURA, S., PECHENIZKIY, M., et al. (eds.) *Handbook of Educational Data Mining.* Boca Raton, FL: CRC Press.

KOKOTT, J. & SOBOTTA, C. 2013. *The distinction between privacy and data protection in the jurisprudence of the CJEU and the ECtHR*. International Data Privacy Law, 3, 4, 222-228.

KOSINSKI, M., STILLWELL, D. & GRAEPEL, T. 2013. *Private traits and attributes are predictable from digital records of human behavior*. Proceedings of the National Academy of Sciences of the USA, 110, 15, 5802–5805.

KUCHLER, H. 2017. *Can Silicon Valley really hack education?* Financial Times (London). Published February 3. Available: https://www.ft.com/content/d629e512-e805-11e6-967b-c88452263daf. [Accessed February 4 2017].

KULIK, J. A. & KULIK, C.-L. C. 1992. *Meta-analytic findings on grouping programs*. Gifted Child Quarterly, 36, 2, 73-77.

KUREKOVÁ, L. M., BEBLAVÝ, M. & THUM, A.-E. 2014. *Using Internet Data to Analyse the Labour Market: A Methodological Enquiry*. Forschungsinstitut zur Zukunft der Arbeit. Published October. Available: http://ftp.iza.org/dp8555.pdf. [Accessed March 4 2015].

LANE, J., STODDEN, V., BENDER, S., et al. (eds.) 2014. *Privacy, Big Data, and the Public Good: Frameworks for Engagement.* Cambridge: Cambridge University Press.

LANEY, D. 2001. *Data Management: Controlling Data Volume, Velocity, and Variety*. Meta Group (Gartner). Published February 6. Available: https://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf. [Accessed January 6 2014].

LARSEN, C., RAND, S., SCHMID, A., et al. (eds.) 2015. *Big Data and the Complexity of Labour Market Policies: New Approaches in Regional and Local Labour Market Monitoring for Reducing Skills Mismatches.* Munich: Rainer Hampp Verlag.

LAVRIJSEN, J. & NICAISE, I. 2015. *New empirical evidence on the effect of educational tracking on social inequalities in reading achievement*. European Educational Research Journal,, 14, 3-4, 206-221.

LAYTON, L. 2015. *Study says standardized testing is overwhelming nation's public schools*. Washington Post. Published October 24. Available: https://www.washingtonpost.com/local/education/study-says-standardized-testing-is-overwhelming-nations-public-schools/2015/10/24/8a22092c-79ae-11e5-a958-d889faf561dc_story.html. [Accessed May 5 2016].

LEU, D. J., EVERETT-CACOPARDO, H., ZAWILINSKI, J., et al. 2013. *New literacies of online reading comprehension*. pp.1-9 *In:* CHAPELLE, C. A. (ed.) *The encyclopedia of applied linguistics.* Hoboken, NJ: Wiley-Blackwell.

LI, T. & LI, N. 2009. *On the tradeoff between privacy and utility in data publishing*. Association for Computing Machinery. Published July. Available: http://dl.acm.org/citation.cfm?id=1557079. [Accessed Decenber 6 2016].

LIPTON, Z. C. 2016. *The Mythos of Model Interpretability*. arXiv.org. Published June. Available: https://arxiv.org/abs/1606.03490. [Accessed December 5 2016].

LITTLEJOHN, A., BEETHAM, H. & MCGILL, L. 2012. *Learning at the digital frontier: a review of digital literacies in theory and practice*. Journal of Computer Assisted Learning, 28, 6, 333-347. Available: DOI: 10.1111/j.1365-2729.2011.00474.x.

LITTLEJOHN, A. & MILLIGAN, C. 2015. *Designing MOOCs for professional learners: Tools and patterns to encourage self-regulated learning*. Open University. Published eLearning Papers, 42, article no. 4. Available: http://oro.open.ac.uk/46385/. [Accessed December 14 2016].

LYNCH, D. J. 2017. *Policing the digital cartels*. Financial Times (London). Published January 8. Available: https://www.ft.com/content/9de9fb80-cd23-11e6-864f-20dcb35cede2. [Accessed January 11 2017].

LYNN, B. C. 2017. *America's Monopolies Are Holding Back the Economy*. The Atlantic. Published February 22. Available: https://www.theatlantic.com/business/archive/2017/02/antimonopoly-big-business/514358/ [Accessed March 19 2017].

MANGAN, K. 2012. *Minnesota Gives Coursera the Boot, Citing a Decades-Old Law*. Chronicle of Higher Education. Published October 18. Available: http://www.chronicle.com/blogs/wiredcampus/minnesota-gives-coursera-the-boot-citing-a-decades-old-law/40542. [Accessed September 6 2015].

MAZUR, E. 2009. *Farewell, Lecture?* Science, 323, 5910, 50-51. Available: http://science.sciencemag.org/content/323/5910/50.

MCINTYRE, E. 2016. *Student privacy concerns hover over big data analytics*. Education Dive. Published January 12. Available: http://www.educationdive.com/news/student-privacy-concerns-hover-over-big-data-analytics/411902/. [Accessed November 3 2016].

MCKINSEY. 2017. *How the world's best-performing schools come out on top*. McKinsey & Company. Published September. Available: http://mckinseyonsociety.com/how-the-worlds-best-performing-schools-come-out-on-top/. [Accessed February 6 2014].

MCLELLAN, C. 2016. *SaaS in 2016: The key trends*. ZDNet.com. Published August 1. Available: http://www.zdnet.com/article/saas-in-2016-the-key-trends/. [Accessed March 19 2017].

MCNELY, B. J., GESTWICKI, P., HILL, J. H., et al. 2012. *Learning analytics for collaborative writing: a prototype and case study*. Proceedings of the 2nd International Conference on Learning Analytics and Knowledge. Published May. Available: http://dl.acm.org/citation.cfm?id=2330654. [Accessed December 2 2015].

MERCER, N., DAWES, L., WEGERIF, R., et al. 2004. *Reasoning as a scientist: ways of helping children to use language to learn science*. British Educational Research Journal, 30, 3, 359-377. Available: http://onlinelibrary.wiley.com/doi/10.1080/01411920410001689689/abstract.

MERCER, N., WEGERIF, R. & DAWES, L. 1999. *Children's talk and the development of reasoning in the classroom*. British Educational Research Journal, 25, 1, 95-111. Available: http://onlinelibrary.wiley.com/doi/10.1080/0141192990250107/full.

MEYER, H.-D. & BENAVOT, A. (eds.) 2013. *PISA, power, and policy: The emergence of global educational governance.* Oxford: Symposium Books Ltd.

MITROS, P., AGARWAL, A. & PARUCHURI, V. 2014. *Assessment in Digital At-scale Learning Environments: MOOCs and technology to advance learning and learning research*. ACSM Ubiquity. Published April. Available: https://doi.org/10.1145/2591795. [Accessed April 6 2015].

MITROS, P. & SUN, F. 2014. *Creating Educational Resources at Scale*. 2014 IEEE 14th International Conference on Advanced Learning Technologies. Available: http://mitros.org/p/papers/pmitros-icalt14.pdf. [Accessed December 3 2016].

MITROS, P. F., AFRIDI, K. K., SUSSMAN, G. J., et al. 2013. *Teaching Electronic Circuits Online: Lessons from MITx's 6.002x on edX*. Proceedings of the IEEE International Symposium on Circuits and Systems (ISCAS), Beijing, China. Published May. Available: Mitros 2013/IEEE 6.002x paper. [Accessed March 2 2017].

MOEDAS, C. 2016. *What new models and tools for measuring science and innovation impact?* European Commission. Published September 26. Available:

http://ec.europa.eu/commission/2014-2019/moedas/announcements/what-new-models-and-tools-measuring-science-and-innovation-impact_en. [Accessed October 12 2016].

MONTELEONE, S. 2017. *Fundamental rights implications of big data*. European Parliamentary Research Service. Published March 10. Available: https://epthinktank.eu/2017/03/10/fundamental-rights-implications-of-big-data-plenary-podcast/. [Accessed March 23 2017].

MOROZOV, E. 2015. *Digital Technologies And The Future Of Data Capitalism*. Social Europe. Published June 23. Available: https://www.socialeurope.eu/2015/06/digital-technologies-and-the-future-of-data-capitalism/. [Accessed January 5 2017].

NARAYANAN, A. & SHMATIKOV, V. 2007. *Robust De-anonymization of Large Datasets: (How to Break Anonymity of the Netflix Prize Dataset)*. arXiv.org. Available: https://arxiv.org/pdf/cs/0610105.pdf. [Accessed December 12 2016].

NAVRACISCS, T. 2016. *Science and policymaking: creating a new dialogue*. European Commission. Published September 29. Available: http://ec.europa.eu/commission/2014-2019/navracsics/announcements/science-and-policymaking-creating-new-dialogue_en. [Accessed October 12 2016].

NCES. 2014. *Table 218.10. Number and internet access of instructional computers and rooms in public schools, by selected school characteristics: Selected years, 1995 through 2008*. National Center for Education Statistics (USA). Available: https://nces.ed.gov/programs/digest/d14/tables/dt14_218.10.asp. [Accessed December 2 2016].

NFCRC. 2017. *"Department of Education" ransomware alert*. National Fraud & Cybercrime Reporting Centre. Published January 4. Available: http://www.actionfraud.police.uk/news/department-of-education-ransomware-alert-jan17. [Accessed January 10 2017].

NIKIFORAKIS, N. & ACAR, G. 2014. *Browse at your own risk*. IEEE Spectrum, 51, 8, 30-35.

NMC. 2016. *The NMC Horizon Report: 2016 Higher Education Edition*. New Media Consortium. Available: http://cdn.nmc.org/media/2016-nmc-horizon-report-he-EN.pdf. [Accessed March 10 2016].

NOONAN, L. 2017. *Deutsche Bank scours social media to find talent*. Financial Times (London). Published January 10. Available: https://www.ft.com/content/84bcb7de-d425-11e6-9341-7393bb2e1b51. [Accessed January 11 2017].

O'CONNOR, S. 2016. *When your boss is an algorithm*. Financial Times (London). Published September 8. Available: http://www.ft.com/cms/s/2/88fdc58e-754f-11e6-b60a-de4532d5ea35.html. [Accessed September 11 2016].

O'NEILL, C. 2016. *Weapons of math destruction: How big data increases inequality and threatens democracy.* New York: Crown Publishing Group.

OCR. 2000. *The Use of Tests as Part of High-Stakes Decision-Making for Students: A Resource Guide for Educators and Policy-Makers*. Office for Civil Rights,, Washington, DC. Published December. Available: https://eric.ed.gov/?id=ED456173. [Accessed February 7 2017].

OECD. 2010. *Education: Governments should expand tertiary studies to boost jobs and tax revenues*. OECD. Available:

http://www.oecd.org/document/52/0,3343,en_21571361_44315115_45925620_1_1_1_1,00.html. [Accessed September 7 2010].

OECD. 2012. *Equity and Quality in Education: Supporting Disadvantaged Students and Schools*. OECD. Available: https://www.oecd.org/education/school/50293148.pdf. [Accessed February 24 2014].

OECD. 2013. *Synergies for Better Learning: An International Perspective on Evaluation and Assessment* OECD. Published April 11. Available: http://www.oecd.org/edu/school/synergies-for-better-learning.htm. [Accessed June 5 2016].

OECD. 2014a. *Education Policy Outlook: Germany*. OECD. Published April. Available: http://www.oecd.org/edu/EDUCATION%20POLICY%20OUTLOOK%20GERMANY_EN.pdf. [Accessed October 5 2016].

OECD. 2014b. *Education Policy Outlook: Portugal*. OECD. Published October. Available: http://www.oecd.org/education/EDUCATION%20POLICY%20OUTLOOK_PORTUGAL_EN.pdf. [Accessed November 4 2016].

OECD. 2015a. *Education Policy Outlook: Canada*. OECD. Published January. Available: http://www.oecd.org/edu/EDUCATION%20POLICY%20OUTLOOK%20CANADA.pdf. [Accessed October 13 2015].

OECD. 2015b. *Education Policy Outlook: Hungary*. OECD. Published November. Available: http://www.oecd.org/edu/Hungary-Profile.pdf. [Accessed June 6 2016].

OECD. 2015c. *Students, Computers and Learning: Making the Connection*. OECD. Published September. Available: http://dx.doi.org/10.1787/9789264239555-en. [Accessed January 6 2016].

OECD. 2016a. *Are there differences in how advantaged and disadvantaged students use the Internet?* OECD. Published July 12. Available: http://www.oecd-ilibrary.org/education/are-there-differences-in-how-advantaged-and-disadvantaged-students-use-the-internet_5jlv8zq6hw43-en. [Accessed July 15 2016].

OECD. 2016b. *Education Policy Outlook: Estonia*. OECD. Published April. Available: http://www.oecd.org/edu/Education-Policy-Outlook-Country-Profile-Estonia.pdf. [Accessed October 6 2016].

OECD. 2016c. *Education Policy Outlook: Slovenia*. OECD. Published April. Available: http://www.oecd.org/edu/Education-Policy-Outlook-Country-Profile-Slovenia.pdf. [Accessed October 7 2016].

OECD. 2016d. *PISA 2015 – Results in Focus*. OECD. Published December 6. Available: http://www.oecd.org/pisa/pisa-2015-results-in-focus.pdf. [Accessed December 6 2016].

OECD. 2016e. *Singapore tops latest OECD PISA global education survey*. OECD. Published December 6. Available: http://www.oecd.org/education/singapore-tops-latest-oecd-pisa-global-education-survey.htm. [Accessed December 6 2016].

OECD. 2016f. *Teachers' ICT and problem-solving skills: Competencies and needs*. OECD. Published April 22. Available: http://www.oecd-ilibrary.org/education/teachers-ict-and-problem-solving-skills_5jm0q1mvzqmq-en. [Accessed May 20 2016].

OHM, P. 2009. *Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization*. UCLA Law Review, 57, 1701-1777. Available: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1450006.

OKUDA, S. 2016. *Defending Student Data from Classrooms to the Cloud: 2016 in Review*. Electronic Frontier Foundation. Published December 23. Available: https://www.eff.org/deeplinks/2016/12/effs-work-student-privacy. [Accessed January 6 2017].

PAMLIN, D. & ARMSTRONG, S. 2015. *12 Risks that threaten human civilisation*. Global Challenges Organisation. Published February. Available: http://globalchallenges.org/publications/globalrisks/about-the-project/. [Accessed February 13 2015].

PEKKARINEN, T., UUSITALO, R. & PEKKALA, S. 2006. *Education policy and intergenerational income mobility: Evidence from the Finnish comprehensive school reform*. Forschungsinstitut zur Zukunft der Arbeit. Published July. Available: http://ftp.iza.org/dp2204.pdf. [Accessed March 2 2017].

PENNEY, J. W. 2016. *Chilling effects: Onlne surveillance and Wikipedia use*. Berkeley Technology Law Journal*,* 31*,* 1*,* 119-181. Available: http://scholarship.law.berkeley.edu/cgi/viewcontent.cgi?article=2104&context=btlj.

PENTLAND, A. 2005. *Socially Aware Computation and Communication*. ACSM Computer*,* 38*,* 3*,* 33-40.

PHILLIPS, J. 2016. *Using Big Data to Improve Business Training*. Trainingindustry.com. Published December 20. Available: https://www.trainingindustry.com/content-development/articles/using-big-data-to-improve-business-training.aspx. [Accessed December 28 2016].

PRINSLOO, P. & SLADE, S. 2014. *Student privacy and institutional accountability in an age of surveillance*. pp.197–214 *In:* MENON, M. E., TERKLA, D. G. & GIBBS, P. (eds.) *Using Data to Improve Higher Education: Research, Policy and Practice. Global Perspectives on Higher Education.* Rotterdam: Sense Publishers.

RAHMAN, A.-U. 2013. *Teacher Assessment and Profiling using Fuzzy Rule based System and Apriori Algorithm*. International Journal of Computer Applications*,* 65*,* 5*,* 22-28. Available: http://research.ijcaonline.org/volume65/number5/pxc3885859.pdf.

REGALADO, A. 2012. *High Stakes in Internet Tracking*. Technology Review. Published June 4. Available: http://www.technologyreview.com/business/40485/. [Accessed June 4 2012].

REID-MARTINEZ, K. & MATHEWS, M. 2015. *Harnessing Data for Better Educational Outcomes*. Center for Digital Education. Available: https://afd34ee8b0806295b5a7-9fbee7de8d51db511b5de86d75069107.ssl.cf1.rackcdn.com/CDE15_Special_Report_Q3_V.PDF [Accessed April 3 2016].

RODEN, L. 2016. *What the Pisa rankings actually tell us about Swedish schools*. The Local (Sweden). Published December 6. Available: http://www.thelocal.se/20161206/what-swedens-improving-school-performance-tells-us-pisa. [Accessed January 31 2017].

RUPPERT, E. S. 2015. *Who Owns Big Data?* Discover Society. Published July 30. Available: http://discoversociety.org/2015/07/30/who-owns-big-data/. [Accessed February 7 2016].

RUUS, K. 2008. *Cyber War I: Estonia Attacked from Russia* The European Institute. Available: http://www.europeaninstitute.org/index.php/component/content/article?id=67:cyber-war-i-estonia-attacked-from-russia. [Accessed January 6 2015].

SABELLA, R. A., PATCHIN, J. W. & HINDUJA, S. 2013. *Cyberbullying myths and realities*. Computers in Human Behavior, 29, 6, 2703-2711.

SALTELLI, A. & GIAMPIETRO, M. 2016. *What is wrong with evidence based policy, and how can it be improved?* Andreasaltelli.eu. Published May. Available: http://www.andreasaltelli.eu/file/repository/FUTURES_Saltelli_Giampietro_6.pdf. [Accessed December 7 2016].

SASS, T. R., ZIMMER, R. W., GILL, B. P., et al. 2016. *Charter High Schools' Effects on Long-Term Attainment and Earnings*. Journal of Policy Analysis and Management, 35, 3, 683–706. Available: http://onlinelibrary.wiley.com/doi/10.1002/pam.21913/abstract.

SCHLEICHER, A. 2015a. *School technology struggles to make an impact*. British Broadcasting Corporation (BBC). Published September 15. Available: http://www.bbc.co.uk/news/business-34174795. [Accessed October 4 2015].

SCHLEICHER, A. 2015b. *Schools for 21st-Century Learners: Strong Leaders, Confident Teachers, Innovative Approaches*. OECD. Published March 12. Available: http://www.oecd-ilibrary.org/education/schools-for-21st-century-learners_9789264231191-en. [Accessed May 13 2015].

SCOTLAND. 2016. *National Improvement Framework for Scottish Education - achieving excellence and equity*. Scottish Government. Published January 6. Available: http://www.gov.scot/Publications/2016/01/8314. [Accessed May 11 2016].

SHACKLOCK, X. 2016. *From Bricks to Clicks - The Potential of Data and Analytics in Higher Education*. Higher Education Commission (UK). Published January. Available: http://www.policyconnect.org.uk/hec/research/report-bricks-clicks-potential-data-and-analytics-higher-education. [Accessed February 10 2016].

SHAMIR, R. 2008. *The age of responsibilization: on market-embedded morality*. Economy and Society, 37, 1, 1-19. Available: http://dx.doi.org/10.1080/03085140701760833.

SHAPIRO, H., ØSTERGAARD, S. F. & ROCCARO, M. 2016. *Education & training 2020: Survey on policies and practices of digital and online learning in Europe*. European Commission. Available: http://bookshop.europa.eu/en/education-training-2020-pbNC0216722/. [Accessed September 13 2016].

SHUTE, V. & VENTURA, M. 2013. *Stealth Assessment: Measuring and Supporting Learning in Video Games*. MIT Press. Available: http://myweb.fsu.edu/vshute/pdf/Stealth_Assessment.pdf. [Accessed June 4 2014].

SIMONITE, T. 2016. *No One Knows How Much Cybercrime Really Costs*. Technology Review. Published January 25. Available: http://www.technologyreview.com/news/545696/no-one-knows-how-much-cybercrime-really-costs/. [Accessed January 25 2016].

SLADE, S. & PRINSLOO, P. 2013. *Learning analytics ethical issues and dilemmas*. American Behavioral Scientist, 57, 10, 1510-1529.

SLADE, S. & PRINSLOO, P. 2015. *Student Perspectives on the Use of their Data: Between Intrusion, Surveillance and Care*. European Journal of Open, Distance and E-learning, 18, 1, online. Available: http://www.eurodl.org/index.php?p=special&sp=articles&inum=6&article=679.

SMC. 2016. *State of the Nation report on social mobility in Great Britain*. Social Mobility Commission (UK). Published November 16. Available:

https://www.gov.uk/government/news/state-of-the-nation-report-on-social-mobility-in-great-britain. [Accessed November 16 2016].

SOLOVE, D. J. 2011. *Nothing to hide. The false trade-off between privacy and security.* New Haven: Yale University Press.

ŠTEFÁNIK, M. 2012. *Internet job search data as a possible source of information on skills demand (with results for Slovak university graduates)*. pp.246-260 *In:* CEDEFOP (ed.) *Building on skills forecasts — Comparing methods and applications.* Luxembourg: European Centre for the Development of Vocational Training.

STRAUMSHEIM, C. 2014. *Massive Closed Online Courses*. Inside Higher Education. Published January 28. Available: https://www.insidehighered.com/news/2014/01/28/state-dept-blocks-access-moocs-countries-economic-sanctions. [Accessed March 4 2014].

SUMMERS, L. 2016. *Men without work*. Financial Times (London). Published September 26. Available: https://www.ft.com/content/9505dbcd-d581-3e90-8578-87239c92db6b. [Accessed September 26 2016].

TAEKKE, J. 2011. *Digital Panopticism and organizational power*. Surveillance and Society, 8, 4, 441-454. Available: http://www.surveillance-and-society.org/ojs/index.php/journal/article/view/luhmann/luhmann.

THILLE, C., SCHNEIDER, E., KIZILCEC, R. F., et al. 2014. *The Future of Data–Enriched Assessment*. Research & Practice in Assessment, 9, 5-21. Available: http://www.rpajournal.com/the-future-of-data-enriched-assessment/.

TSORMPATZOUDI, P., BERENDT, B. & COUDERT, F. 2016. *Privacy by Design: From research and policy to practice - the challenge of multi-disciplinarity*. pp.199-212 *Privacy Technologies and Policy. Third Annual Privacy Forum, APF 2015. Luxembourg, Luxembourg, October 7-8, 2015. Revised Selected Papers.* Berlin: Springer.

TURKLE, S. 2011. *Alone Together: Why We Expect More from Technology and Less from each Other.* New York: Basic Books.

TYMMS, P. 2016. *Whose fault is PISA?* University of Durham. Published December 16. Available: https://www.dur.ac.uk/research/news/thoughtleadership/?itemno=30008. [Accessed December 20 2016].

ULD. 2012. *Personenbezogene Datenverarbeitung der Schulen*. Unabhängiges Landeszentrum für Datenschutz Schleswig-Holstein. Published April 17. Available: https://www.datenschutzzentrum.de/schule/20120417-strategiepapier-schulen.html. [Accessed December 5 2015].

UN. 1989. *Convention on the Rights of the Child*. United Nations. Published November 20. Available: https://treaties.un.org/doc/Treaties/1990/09/19900902%2003-14%20AM/Ch_IV_11p.pdf. [Accessed February 24 2013].

UNESCO. 1996. *Learning: The Treasure Within*. UNESCO. Available: http://unesdoc.unesco.org/images/0010/001095/109590eo.pdf. [Accessed February 28 2017].

UNESCO. 2015. *Rethinking Education: Towards a global common good?* UNESCO. Available: http://unesdoc.unesco.org/images/0023/002325/232555e.pdf. [Accessed March 1 2017].

UNESCO. 2016a. *Designing effective monitoring and evaluation of education systems for 2030: A global synthesis of policies and practices*. UNESCO. Published January.

Available: http://www.unesco.org/new/fileadmin/MULTIMEDIA/HQ/ED/pdf/me-report.pdf. [Accessed June 4 2016].

UNESCO. 2016b. *Methodology of national education accounts*. UNESCO. Available: http://www.uis.unesco.org/Education/Documents/nea-visual-results-report/NEA_Methodology.pdf. [Accessed January 6 2017].

UNESCO. 2016c. *Target 4.5 – What is at stake for monitoring progress on equity in education?* UNESCO. Published October 18. Available: https://gemreportunesco.wordpress.com/2016/10/18/target-4-5-what-is-at-stake-for-monitoring-progress-on-equity-in-education/. [Accessed January 3 2017].

UNICEF. 2010. *Equity and Inclusion in Education: A guide to support education sector plan preparation, revision, and appraisal*. United Nations International Children's Emergency Fund. Published April. Available: https://www.unicef.org/education/files/Equity_and_Inclusion_Guide.pdf. [Accessed March 4 2014].

UNICEF. 2013. *Education Equity Now! A regional analysis of the situation of out of school children in Central and Eastern Europe and the Commonwealth of Independent States*. United Nations International Children's Emergency Fund. Available: http://education-equity.org/wp-content/uploads/2013/12/Unicef_Global-Initiative-on-Out-of-School-Children_Summary_En_FINAL.pdf. [Accessed March 2 2016].

VAN RIJMENAM, M. 2014. *Four Ways Big Data Will Revolutionize Education*. Datafloq.com. Published April 29. Available: https://datafloq.com/read/big-data-will-revolutionize-learning/206. [Accessed December 4 2016].

VANLEHN, K. 2011. *The Relative Effectiveness of Human Tutoring, Intelligent Tutoring Systems, and Other Tutoring Systems*. Educational Psychologist, 46, 4, 197-221. Available: http://dx.doi.org/10.1080/00461520.2011.611369.

VINCENT-LANCRIN, S. & GONZÁLEZ-SANCHO, C. 2014. *Fostering Innovation and Improvement in Education: the Contribution of Longitudinal Information Systems*. OECD. Published July. Available: http://www.oecd.org/edu/ceri/workshop-fostering-innovation-improvement-education-longitudinal-information-systems.htm. [Accessed June 6 2016].

VON STUMM, S. & PLOMIN, R. 2015. *Socioeconomic status and the growth of intelligence from infancy through adolescence*. Intelligence, 28, 30-36. Available: http://dx.doi.org/10.1016/j.intell.2014.10.002.

VOOGT, J., ERSTAD, O., DEDE, C., et al. 2013. *Challenges to learning and schooling in the digital networked world of the 21st century*. Journal of Computer Assisted Learning, 29, 5, 403-413.

WADE, M. R. & PARENT, M. 2002. *Relationships Between Job Skills and Performance: A Study of Webmasters*. Journal of Management Information Systems, 18, 3, 71-96. Available: http://www.tandfonline.com/doi/abs/10.1080/07421222.2002.11045694

WAKEFIELD, J. 2017. *Is your child a cyberbully and if so, what should you do?* British Broadcasting Corporation (BBC). Published January 6. Available: http://www.bbc.co.uk/news/technology-38529437. [Accessed January 9 2017].

WARD, M. 2017. *CES 2017: New routers defend smart homes against hacks*. British Broadcasting Corporation (BBC). Published January 7. Available: http://www.bbc.com/news/technology-38415067. [Accessed January 7 2017].

WARRELL, H. 2015. *Students under surveillance*. Financial Times (London). Published July 24. Available: http://www.ft.com/cms/s/2/634624c6-312b-11e5-91ac-a5e17d9b4cff.html. [Accessed July 29 2015].

WASTIAU, P., BLAMIRE, R., KEARNEY, C., et al. 2013. *The Use of ICT in Education: a survey of schools in Europe*. European Journal of Education, 48, 1, 11-27.

WATKINS, A. & EBERSOLD, S. 2016. *Efficiency, Effectiveness and Equity within Inclusive Education Systems*. pp.229-253 *In:* WATKINS, A. & MEIJER, C. (eds.) *Implementing Inclusive Education: Issues in Bridging the Policy-Practice Gap (International Perspectives on Inclusive Education, Volume 8)* Bingley, UK: Emerald Group Publishing Ltd.

WATTERS, A. 2015. *Ed-Tech's Inequalities*. Hackeducation.com. Published April 8. Available: http://hackeducation.com/2015/04/08/inequalities. [Accessed March 1 2017].

WEF. 2014. *Matching Skills and Labour Market Needs: Building Social Partnerships for better skills and better jobs*. World Economic Forum. Published February 3. Available: https://www.weforum.org/reports/matching-skills-and-labour-market-needs-building-social-partnerships-better-skills-and-better-jobs. [Accessed June 4 2014].

WELCH, M. J. 2011. *Eight-year study and charter legitimacy*. Journal of Education, 191, 2, 55-65.

WESTIN, A. F. 1970. *Privacy and freedom.* New York: Atheneum.

WHITTHAUS, G., INAMORATO DOS SANTOS, A., CHILDS, M., et al. 2016. *Validation of Non-formal MOOC-based Learning: An Analysis of Assessment and Recognition Practices in Europe (OpenCred)*. Joint Research Centre (European Commission). Available: https://ec.europa.eu/jrc/en/publication/eur-scientific-and-technical-research-reports/validation-non-formal-mooc-based-learning-analysis-assessment-and-recognition-practices. [Accessed March 1 2016].

WILLIAMSON, B. 2015a. *Cognitive computing and data analytics in the classroom*. British Sociological Association. Published April. Available: http://www.academia.edu/11968853/Cognitive_computing_and_data_analytics_in_the_classroom. [Accessed June 4 2016].

WILLIAMSON, B. 2015b. *Educating the smart city: Schooling smart citizens through computational urbanism*. Big Data and Society, July-December, 1013. Available: http://bds.sagepub.com/content/2/2/2053951715617783.

WILLIAMSON, B. 2016. *Critical questions for big data in education*. University of Stirling. Published June 2. Available: https://codeactsineducation.wordpress.com/2016/06/02/critical-questions-for-big-data-in-education/. [Accessed November 21 2016].

WILLIAMSON, B. & FACER, K. 2004. *More than'just a game': the implications for schools of children's computer games communities*. Education, Communication & Information, 4, 2-3, 255-270.

WIRED. 2017. *The next generation of brain controlled prosthetics will be for everyone*. Wired.com. Published February. Available: http://www.wired.co.uk/article/how-your-brain-will-control-the-next-generation-of-prosthetics. [Accessed February 23 2017].

WOOD, P. 2016. *The British data-crunchers who say they helped Donald Trump to win*. Spectator (London). Published December 3. Available:

http://www.spectator.co.uk/2016/12/the-british-data-crunchers-who-say-they-helped-donald-trump-to-win/. [Accessed December 3 2016].

WÖßMANN, L. & SCHÜTZ, G. 2006. *Efficiency and Equity in European Education and Training Systems* European Commission. Published April 26. Available: http://www.erisee.org/downloads/equity/eenee.pdf. [Accessed March 3 2014].

WOWCZKO, I. 2015. *Skills and Vacancy Analysis with Data Mining Techniques*. Informatics*,* 2*,* 31-49. Available: www.mdpi.com/2227-9709/2/4/31/pdf

WRIGHT, B. D. & STONE, M. H. 1979. *Best Test Design: Rasch Measurement.* Chicago: MESA Press.

WRIGHT, D. & KREIS, R. (eds.) 2014. *Surveillance in Europe.* London: Routledge.

YOON, D. & MITROS, P. 2015. *Multi-Modal Peer Discussion with RichReview on edX*. ACM User Interface and Technology Symposium. Available: http://mitros.org/p/papers/2015_uist.pdf. [Accessed December 25 2016].

ZEIDE, E. 2016. *Student Privacy Principles for the Age of Big Data: Moving beyond FERPA and FIPPs*. Drexel Law Review*,* 8*,* 339-396.

ZIEGELE, F. & VAN VUGHT, F. 2017. *Towards a fairer and more robust university ranking*. University World News. Published January 8. Available: http://www.universityworldnews.com/article.php?story=20161221132830722. [Accessed January 10 2017].

Annex A: The Commission brief

Pupils' and students' privacy
Already experiments with classrooms with a variety of regular and infrared cameras, microphones, wearable tracking devices and laptop/tablet tracking software are implemented. Such data harvesting will potentially allow for searching for patterns in each student's engagement level, moods, use of classroom resources, social habits, language and vocabulary use, attention span and academic performance in order to gain insights in optimising learning processes and environments. These developments are for many both controversial and disturbing, and call for a rich debate on privacy matters in educational systems.

Educational efficiency and equity
Efficiency is a multidimensional concept that also ties into national political discourses on distribution of funding. Strong effects on the overall equity performance of educational systems are identified, but effective policy measures calls for new and better sources of information. The possibility to couple rich, granular and live data on pupils' academic performance with data on a variety of context parameters, such as general administrative data, fiscal spending data for schools, teachers' qualifications, indicators on the socio-economic environment for the school and aggregated academic performance represents only the beginning of data harvesting but will have profound effects on the knowledge of educational efficiency and equity.

Student tracking
When artificial intelligence is used over time on the mix of pupil's performance data and data as in the examples above, very strong predictive power in the analytical models of pupils' academic and skills developments are expected. An apparent trait would be that ability tracking of students might become very accurate at a much earlier age than today, and such information might feed into national level monitoring and policy development of educational systems.

Assessment
Technology offers enhanced question types and measurement procedures, allowing for testing more dimensions of already established competency frameworks and the measurement of complex competences, also including non-cognitive skills. This area of big data usage in education points towards increasingly rich large-scale databases of assessment results, with aggregation at different levels (local, national, international).

Skills forecasting
Targeting the alignment of labour market skills demands and educational systems' candidate output will be influenced by use of big data. By using national statistics on school and study trajectories and employment history, emerging skills demand can be detected more precisely than only by relying on business sectors' self-reports. Big data is also already used to automatically scan and categorise online job advertisements in order to provide a real-time snapshot of skills demands.

Annex B: The Authors

Professor Bettina Berendt is a professor in the research group "Declarative Languages and Artificial Intelligence" (DTAI) at the Department of Computer Science of KU Leuven. Her research focuses on Web Mining and its uses and implications. Methodologically, she combines aspects of Web content, Web usage and Web structure mining with methods from the social and behavioural sciences relevant to the respective research questions and applications. The Web materials and platforms include Social Media such as microblogging or social networking sites, as well as (mainstream or other) news sites and the relation between these various channels of information. Research questions include privacy, the public and the private and the role of media in them, information literacy, and how data mining can be developed and deployed in a user-centric fashion for user empowerment in these areas.

Philippe Kern is the Managing Director and Founder of KEA, he has 25 years of experience in culture policies, creative industries at international level with expertise in research, mapping, cluster development, valorisation of the value-chain and creative entrepreneurship. He has been pivotal in the design and delivery of translational exchange and learning activities, thematic expertise in local economic development; research, innovation and knowledge economy; Entrepreneurship and competitive SMEs; Arts and Culture. Philippe is experienced in embedding big data in evaluation, monitoring and statistical frameworks. As part of a feasibility study on data collection and analysis in the cultural and creative sectors in the EU, the proposed statistical framework put forward recommendations to combine big data (web scraping, social media monitoring, cooperation with collecting societies and business associations) with existing data collection mechanisms. Philippe's past projects involved big data management as well as strategies to facilitate user engagement through big data, as he was managing strategic advice assignments (for example in a mission on the development of living labs in Wallonia), or communication activities (Creative Tracks and Sparks projects).

Professor Allison Littlejohn is the Academic Director for Digital Innovation and Chair of Learning Technology at the Institute of Educational Technology at The Open University. She has worked throughout her career in the area of learning innovation, education, technology, knowledge creation and academic-business partnerships. She has worked with multinational companies, including Shell, BP International and Conoco-Philips. During 2008 – 2010 she was a Senior Researcher for Royal Dutch Shell where she led a university-industry partnership in technology enhanced learning. She specialises in education and learning, exploring how expertise development can be supported and enhanced by information and communication technologies, including social media. She has produced many publications in this area, including Learning in open networks for work, life and education in 2014.

Piotr Mitros is the Chief Scientist of edX where he leads research and development initiatives. Mitros is a frequent conference keynote speaker or panellist on disruption in education, assessment, learning analytics, educational datamining, open educational resources and crowdsourcing in education. His observations of university systems around the world inspired him to find innovative ways to dramatically increase both the quality of and access to education. Recent publications include "Big Data Analysis in Higher Education: Promises and Pitfalls" and "Data-Intensive Research in Education: Current Work and Next Steps".

Xanthe Shacklock is an experienced policy professional with a deep understanding of the educational and skills system in the UK, especially higher education policy. She has a strong background in researching and producing long-form policy reports. Xanthe led on the Higher Education Commission's fourth research inquiry, wrote and produced the final inquiry report, "From Bricks to Clicks: The Potential of Data and Analytics in Higher

Education." This included setting terms of reference and circulating a call for evidence, organising evidence sessions, conducting interviews with key stakeholders and writing the final report.

HOW TO OBTAIN EU PUBLICATIONS

Free publications:

- one copy:
  via EU Bookshop (http://bookshop.europa.eu);
- more than one copy or posters/maps:
  from the European Union's representations (http://ec.europa.eu/represent_en.htm);
  from the delegations in non-EU countries (http://eeas.europa.eu/delegations/index_en.htm);
  by contacting the Europe Direct service (http://europa.eu/europedirect/index_en.htm) or
  calling 00 800 6 7 8 9 10 11 (freephone number from anywhere in the EU) (*).

  (*) The information given is free, as are most calls (though some operators, phone boxes or hotels may charge you).

Priced publications:

- via EU Bookshop (http://bookshop.europa.eu).

NC-04-17-442-EN-N

Publications Office