10 International trends in educational research output: a bibliometric approach

Stéphan Vincent-Lancrin, OECD

Gwénaël Jacotin, OECD

This chapter proposes a method to measure educational research based on academic publications. After presenting a semantic approach to identifying educational research output, it shows how the production and geographic distribution of the educational research output evolved over the past 30 years, how different disciplines contributed to this output and looks at other characteristics of educational research such as its collaborative nature or the balance between qualitative and quantitative publications.

Introduction

Bibliometrics is commonly used to measure the quantity and distribution of research across countries and topics, and sometimes the focus, collaborative nature and citation impact of countries' research (OECD, 2017[1])[1]. Countries have performed such bibliometric analysis of their educational research, either occasionally: Australia (Phelan, Anderson and Bourke, 2000[2]); Chile: (Brunner and Salazar, 2009[3]); or on a regular basis, e.g. Norway (Gunnes and Rørstad, 2015[4]). In some cases, a bibliometric approach has also been used to assess the distribution of educational research across countries (Sezgin, Orbay and Orbay, 2022[5]), the performance of educational research (Diem and Wolter, 2013[6]), the impact of COVID on educational research (Cretu and Ho, 2023[7]), etc.

This chapter explores the extent to which this method can be applied to the OECD educational research output, and, more broadly, to the global output of educational research. The main contribution of the chapter is to propose a methodology to define educational research through a semantic search rather than databases' pre-identified "education" categories. Indeed, educational research can be performed by researchers in different disciplines, and is a "topic of interest" as much as a "discipline". Limiting the educational research output to pre-categorised "education" categories may lead to missing a significant share of educational research.

The second objective of the chapter is to describe and analyse the evolution of the educational research output over time, its distribution across countries, but also a comparison with the overall research output or research outputs in other disciplines. We also show how educational research is distributed by discipline, and show that further semantic search could allow one to identify other characteristics of interest, for example how much of educational research is quantitative or qualitative. The conclusion points to what a regular data collection on educational research could look like.

Data and methodology

This section presents the main aspects of the bibliometric methodology used in this study. Annex 10.A presents it in more details. The main point is that we identify the educational research output (and thus the universe of our analysis) through a semantic search in the title and abstracts of the research papers.

The data source for this study is the open LENS bibliometric application about academic outputs. The application inventories scholar works from the following publication databases: OpenAlex (formerly Microsoft Academic Graph [MAG]), Crossref, PubMed, Core and PubMed Central. LENS provides information about the catalogued research output such as title, abstract, author's affiliations, year of publication, type of publication, discipline, etc. The databases cover over 240 million scientific documents, but this study focuses on the outputs included in the "analytics set" of the application (45% of the total database, that is 108 million documents): author information is indeed available for those documents, which include journal articles, book chapters, conference proceedings, conference proceeding articles, books, letters or reviews. Among them, over 82% are journal articles and 10%, book chapters.

The methodology to identify educational research articles is based on the presence of certain words or strings of letters (such as "educat", "student" or "teach") in the article title and/or abstract[2]. The choice of the words was done through an iterative process, with a manual verification on random sub-samples that the corresponding scientific output could be qualified as educational research. The method is a mix of trial and error based on the gradual elimination of "noise" within the different sub-samples of the corresponding universe, trying to find a limited combination of words that would not be too large and include too many non-educational articles or too narrow (and possibly miss too many). Given the size of the samples, only a probabilistic approach is possible. It arguably misses some academic educational research papers, includes some that are not "educational research", but provides a good estimate of the total educational research output, and, perhaps more importantly, of its evolution over time (assuming that the level of noise

and mistakes remains the same over time). For this reason, most of the analysis is presented in terms of evolution of percentages and shares rather than as the number of outputs.

This method allowed us to define a relevant corpus of educational research output comprised of 2.6 million documents, an overwhelming majority being research articles. Educational research thus represents about 1% of the total research output – a share that (coincidentally) roughly corresponds to the 1% of public research funding it receives in the OECD area (Vincent-Lancrin, 2023[8]). Out of those documents, only 52% have an identified author and affiliation, which limits our useable dataset to 1.4 million documents. This is still a reasonable amount of information, acknowledging that the lack of useable information also concerns other research subjects (and databases). The assumption is that the remaining sample is representative of the 48% of documents for which we do not have sufficient information.

Trends in educational research

This section presents some of the big trends of educational research in terms of publication output. It covers the quantity of educational research output over time, its geographic distribution as well as some of its characteristics (discipline, collaboration, quantitative vs qualitative).
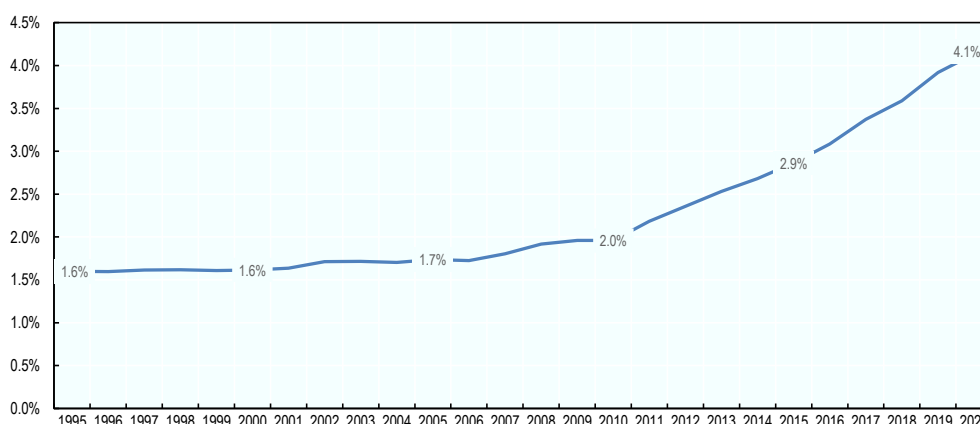
An increase in the educational research output

Figure 10.1. Educational research outputs in the world, 1995-2020
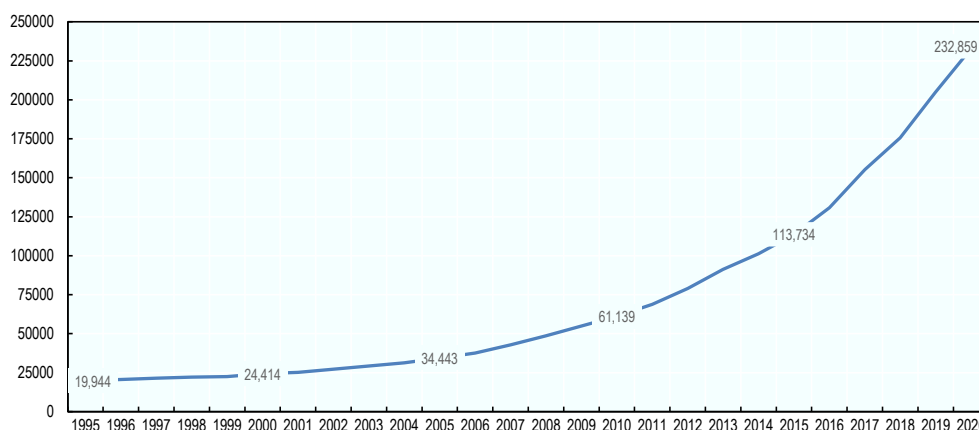
Note: Authors' calculations adapted from the LENS analytics set

StatLink ⧉ https://stat.link/f6jyol

Figure 10.2. Share of educational research outputs in the total research output, 1995-2020

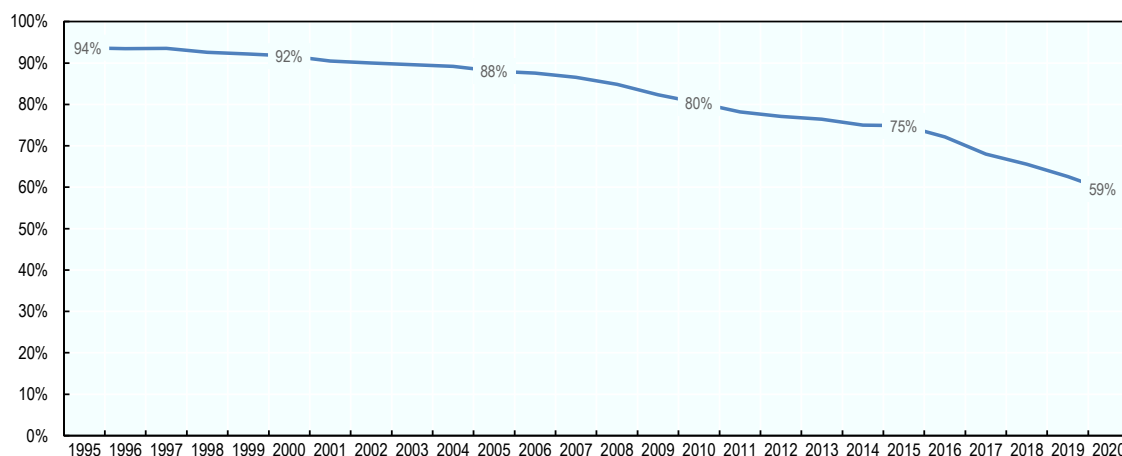Source: Authors' calculations adapted from the LENS analytics set

StatLink 📊 https://stat.link/tr4vgc

Figure 10.1 and Figure 10.2 show that the world educational research output has significantly increased, both in absolute and in relative terms (compared to the total research output). After a relatively stable production between 1995 and 2010, it has skyrocketed, going from 61 000 academic documents in 2010 to 233 000 in 2020. As a share of the total global research production, educational research went up from 1.6% to 4.1% between 1995 and 2020, showing that the production of educational research has grown more quickly than the overall research production. Part of this acceleration comes from the incentives provided to academics to publish more, whatever the domain, but as the trend in educational research outpaces the increase in the overall research output, it corresponds to a real increase in the interest of researchers and possibly research funders in education as a topic, possibly with the hope that research will contribute to educational improvement (OECD, 2007[9]; OECD, 2022[10]).

A declining share of OECD countries' output in the world output

Figure 10.3 shows that OECD countries produce most of the educational research in the world. Given the rapid increased of non-OECD countries' output, the share of the OECD area in the total educational research output has decreased over time, especially since 2015. In 2020, OECD countries produced 59% of the educational research publications, against 80% in 2010 and 94% in 1995. The OECD country output has continued to grow, but countries such as China, Brazil and Indonesia have significantly increased their output from a very low starting point during the same period. Educational research is becoming more important and distributed globally, and not just within the OECD, which opens new possibilities for international collaboration and peer learning. As educational improvement is also situated in a certain country context and the gained knowledge is not always transferable across countries, the local production of educational research provides more opportunities for the improvement of education worldwide.

Figure 10.3. Share of educational research outputs of OECD countries in total educational research output, 1995-2020
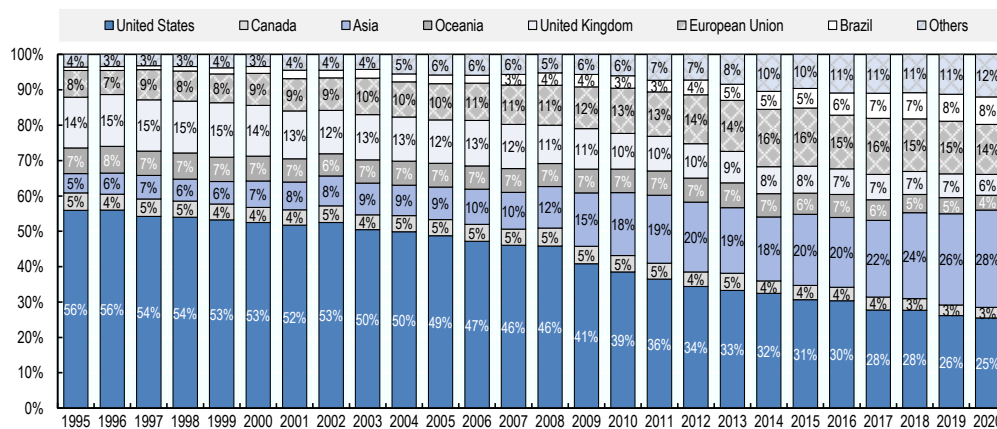

Source: Authors' calculations adapted from the LENS analytics set

Figure 10.4 presents the distribution of educational research in the world by major countries or areas of production as a share of the total research output.[3] The United States has remained by far the largest country producing educational research outputs within the OECD and in the world, with a 5-fold increase of its outputs between 1995 and 2020. However, its share in the world production has steadily decreased as a result of the growth of other countries' output. Since 2008, educational research publications from Asian countries, notably China and Indonesia, have skyrocketed and reached US levels. The production of the European Union and Brazil has also increased significantly – while the educational research of Canada, the United Kingdom or Oceania has also doubled, but increased much less than in most other countries. That being said, as shown in Figure 10.4, the United Kingdom is one of the largest educational research producer in the OECD and in the world, accounting for about 6% of education research articles in 2020, and one of the few countries that concentrate about 3% or more of the educational research output in the world with the United States (26%), Brazil (8%), Indonesia (6%), China (5%), Australia (3%), Canada (3%) and Spain (3%).

Figure 10.4. Share of countries/regions' output in total educational research output, 1995-2020

Source: Authors' calculations adapted from the LENS analytics set

Table 10.1. Country share of the world's educational research output, 1990-2020

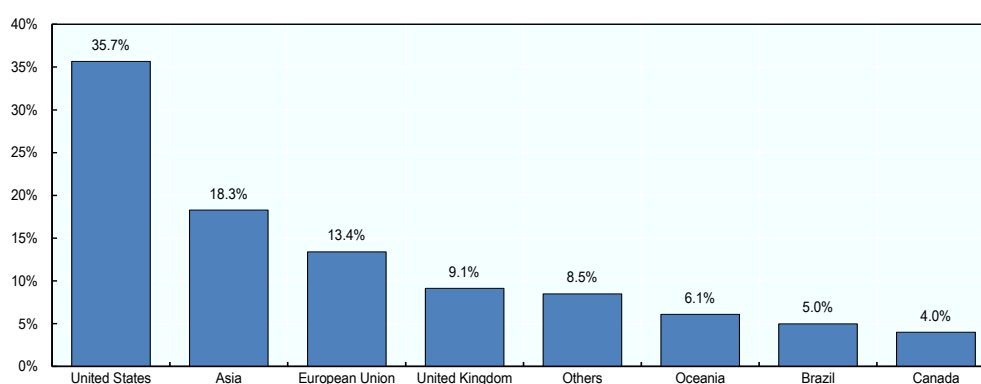|	1990	1995	2000	2005	2010	2015	2020
Australia	4.4%	6.6%	6.1%	5.6%	5.4%	4.9%	3.5%
Austria	0.1%	0.1%	0.1%	0.2%	0.2%	0.3%	0.3%
Belgium	0.2%	0.3%	0.5%	0.4%	0.5%	0.5%	0.4%
Canada	4.4%	5.0%	4.2%	4.6%	4.6%	4.0%	3.0%
Chile	0.3%	0.3%	0.2%	0.3%	0.4%	0.5%	0.4%
Colombia	0.1%	0.1%	0.1%	0.3%	0.5%	1.0%	0.6%
Costa Rica	0.0%	0.0%	0.0%	0.0%	0.1%	0.1%	0.1%
Czech Republic	0.2%	0.1%	0.1%	0.1%	0.2%	0.6%	0.3%
Denmark	0.3%	0.2%	0.2%	0.3%	0.3%	0.5%	0.3%
Estonia	0.0%	0.0%	0.0%	0.1%	0.1%	0.2%	0.1%
Finland	0.4%	0.5%	0.5%	0.5%	0.6%	0.7%	0.6%
France	0.4%	0.3%	0.4%	0.5%	0.4%	0.3%	0.3%
Germany	1.7%	1.2%	1.7%	1.4%	1.8%	1.9%	2.0%
Greece	0.0%	0.2%	0.4%	0.6%	0.7%	0.6%	0.6%
Hungary	0.0%	0.1%	0.1%	0.1%	0.1%	0.2%	0.2%
Iceland	0.0%	0.0%	0.0%	0.0%	0.1%	0.1%	0.0%
Ireland	0.2%	0.3%	0.4%	0.6%	0.8%	0.6%	0.5%
Israel	1.3%	1.2%	1.3%	1.3%	0.7%	0.6%	0.5%
Italy	1.1%	0.6%	0.6%	0.7%	0.8%	1.1%	1.0%
Japan	1.3%	1.2%	1.4%	1.3%	1.6%	1.6%	1.3%
Korea	0.1%	0.2%	0.4%	0.5%	1.0%	2.2%	0.7%
Latvia	0.1%	0.0%	0.1%	0.1%	0.1%	0.3%	0.1%
Lithuania	0.0%	0.0%	0.0%	0.1%	0.1%	0.4%	0.2%
Luxembourg	0.0%	0.0%	0.0%	0.0%	0.0%	0.1%	0.0%
Mexico	0.3%	0.2%	0.2%	0.3%	0.4%	0.9%	1.1%
Netherlands	1.1%	1.2%	1.2%	1.2%	1.2%	1.0%	0.6%
New Zealand	0.9%	0.6%	0.9%	1.2%	1.1%	1.0%	0.6%
Norway	0.4%	0.4%	0.3%	0.4%	0.5%	0.5%	0.6%
Poland	0.1%	0.1%	0.1%	0.2%	0.3%	0.7%	0.9%

	1990	1995	2000	2005	2010	2015	2020
Portugal	0.1%	0.1%	0.2%	0.2%	0.4%	0.6%	0.6%
Slovak Republic	0.0%	0.1%	0.1%	0.1%	0.0%	0.3%	0.2%
Slovenia	0.0%	0.0%	0.2%	0.2%	0.2%	0.2%	0.2%
Spain	0.7%	0.8%	1.1%	1.4%	2.3%	3.3%	3.1%
Sweden	0.5%	0.8%	0.9%	1.1%	1.1%	1.2%	1.0%
Switzerland	0.4%	0.3%	0.3%	0.3%	0.4%	0.4%	0.3%
Turkey	0.2%	0.2%	0.3%	0.9%	2.6%	3.0%	1.6%
United Kingdom	11.9%	14.4%	14.4%	12.3%	10.1%	7.6%	5.9%
United States	60.6%	55.9%	52.5%	48.7%	38.5%	30.6%	25.5%
OECD total	**93.8%**	**93.7%**	**91.7%**	**88.0%**	**80.4%**	**74.8%**	**59.1%**
Argentina	0.1%	0.1%	0.1%	0.2%	0.2%	0.4%	0.5%
Brazil	0.6%	1.0%	2.0%	2.5%	3.5%	5.5%	7.8%
Bulgaria	0.1%	0.1%	0.0%	0.0%	0.0%	0.1%	0.2%
Croatia	0.0%	0.2%	0.1%	0.1%	0.1%	0.2%	0.2%
Peru	0.0%	0.0%	0.0%	0.0%	0.1%	0.1%	0.3%
Romania	0.0%	0.0%	0.0%	0.0%	0.3%	0.9%	0.4%
China	0.4%	0.7%	1.4%	1.7%	6.3%	3.2%	5.3%
India	0.5%	0.4%	0.5%	0.7%	1.0%	1.6%	2.0%
Indonesia	0.1%	0.0%	0.0%	0.0%	0.1%	1.0%	6.0%
South Africa	0.9%	0.8%	0.8%	1.5%	1.5%	2.4%	2.1%

Source: Authors' calculations adapted from the LENS analytics set

Figure 10.5 presents the cumulative stock of educational research output from 1995 to 2020 by country or region of origin. As current research is likely to build on what has been produced in the past decades (and not just the past years), this gives us an idea of the possible geographical influence of research from different countries/regions (assuming all papers have the same chance to have an influence). Ideally, this should be supplemented by an analysis of the citation impact of research from different regions, which was not possible in the framework of this study. The United States and Canada have produced about 40% of the "legacy" research output, the Asia-Pacific region, about 24%, and EU countries and the United Kingdom, about 22%.

Figure 10.5. Cumulative stock of educational research by country/region of production, 1995-2020

Source: Authors' calculations adapted from the LENS analytics set

StatLink https://stat.link/k1bje8

While the quantity of output produced every year matters to give an idea of where production is most valued, it does not provide an indicator of the quality of the produced educational research. One could even argue that a smaller but higher quality output is more helpful to improve the quality of education than a large output with mixed quality research. On the other hand, educational research is developed by producing medium level capacity outputs and gradually building a culture of research excellence.

The increase in China's output mirrors (to a much lesser extent) what can be observed in other scientific domains and other papers using different databases (Sezgin, Orbay and Orbay, 2022[5]). The increase in the shares of Brazil and Indonesia as large producers of educational output should be taken with more caution as this is not a trend that was documented elsewhere. It might come from the database used and an uneven distribution of papers with identified authors and abstracts. Another explanation could also be measurement error that would entail more "false positives" (educational research documents that should not be included in this category) in these countries than in others. A quick review of their recent production seems to indicate that most educational research outputs are produced in medicine, science and computing – areas where articles are much shorter than the typical social science documents. This "specialisation" could also explain the more rapid growth compared to other countries. This analysis is beyond the scope of this study though.

### *How much different fields of study contribute to educational research*

Identifying educational research articles through a semantic search allows one to identify papers produced in more disciplines than "education sciences". This is important as education is as much a subject as it is a discipline. While in some countries it has become a discipline (with its education schools or faculties), in others it is still a "subject" dealt with by a number of other actors: economists, sociologists, political scientists, but also scientists or computer scientists who do not necessarily identify as "educational researchers" but as scientists from another domain working on education, often doing "discipline-based education research".

In the LENS database, all articles are associated with a disciplinary subject such as "education", "general medicine", "developmental and educational psychology", "public health, environmental and occupational health", "sociology and political science", "linguistics and language", etc. The taxonomy includes 332 different subjects, which we reclassified according to the international taxonomy of "fields of science": natural sciences (1), engineering and technology (2), medical and health sciences (3), agricultural sciences (4), social sciences (5), and humanities (6). In the case of social sciences, where we expected to find a significant amount of education research outputs, we also went to the next level of the classification: psychology (5.1), economics and business (5.2), educational sciences (5.3), sociology (5.4), law (5.5), political science (5.6), social and economic geography (5.7), media and communications (5.8), and other social sciences (5.9).

Figure 10.6 shows the distribution of educational research across fields of science. Social sciences accounted for 40% of education research in 2020, a small decrease from 45% in 1995, while humanities remained stable around 10%. What we typically have in mind when thinking of educational research thus represent about half of it. A significant amount of research is also carried out in the fields in which countries spend more of their public research budget: medical and health sciences (35% of the output in 2020), natural sciences (11%) and engineering and technology (4%).

Figure 10.6. Distribution of educational research by field of science, 1995-2020
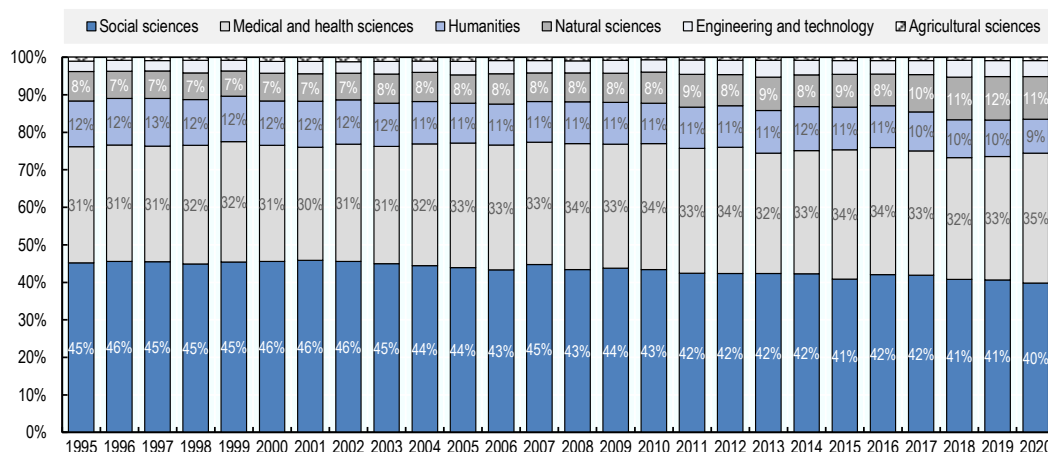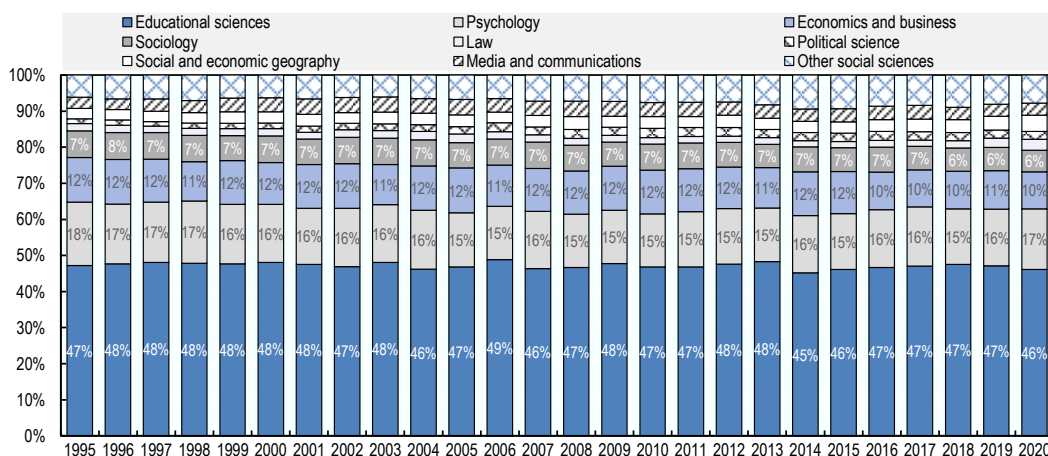

Source: Authors' calculations adapted from the LENS analytics set

StatLink ⟱ https://stat.link/jhz9io

In 2020, articles classified as educational sciences represented 18% of the total educational research output, psychology, 7%, economics, 4%, sociology, 2%, and the other social science fields, the remaining 8%. Here again, it is noteworthy that the distribution remained stable over time. Figure 10.7 presents the same information by zeroing in on the social science output of educational research: educational sciences and psychology represent about 60% of the social science educational research output.

Figure 10.7. Share of educational sciences, psychology and other sub-fields of social sciences in the total educational research output of social sciences, 1995-2020

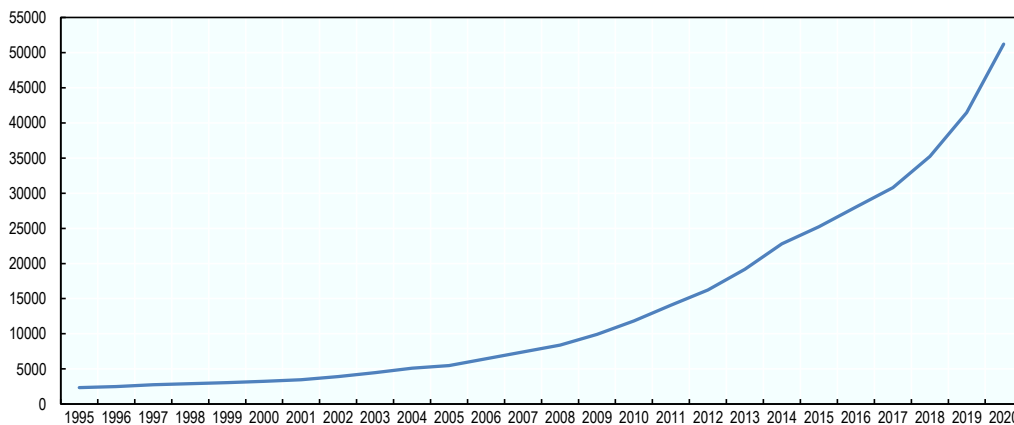Source: Authors' calculations adapted from the LENS analytics set

StatLink ⟱ https://stat.link/riv1qh

An increasing amount of quantitative educational research outputs

Discussions on educational research in the past decades included considerations about the balance between different types of research, what counts as "evidence" and how research is used (OECD, 2007[9]; OECD, 2022[10]). Some have lamented that educational research was based on weak theoretical bases that did not give more room to neuroscience. Others have complained that it was too estranged from causal inference, with debates about different ways to get close to such causal inference (Schneider et al., 2007[11]). Sometimes, the debate turns around the share between qualitative and quantitative research, with the idea that quantitative empirical research which can generalise to an entire education system is still not frequent enough, partly because of a lack of data collected by governments or made available to researchers, partly because of the epistemic traditions of educational science, which was initially anchored in humanities, developmental psychology and pedagogical research. One can assume that the surge of data collected within countries as well as the shift towards exploring causal inference in social science may have led to an increase in quantitative educational research papers.

We have tried to capture the quantitative nature of different articles through another semantic. We identified the educational research papers with the following words in their abstract: "data", "sample", "statistic", "control group" as well as "quantitativ" and "estimat". The same iterative method including a manual check of a random sample showed that the technique was reliable to identify papers using a quantitative methodology – acknowledging that this is just an estimate. Again, trend values thus represent the most important aspect of the analysis. Since 2009, quantitative educational research papers have increased 5-fold (assuming that the trend was the same for articles that missed an abstract and could not be analysed, see Figure 10.8). Figure 10.9 shows that the percentage of quantitative papers in total educational research output has almost doubled since 1995 and significantly increased since 2009 to represent about 36% of the total research output in 2020.

Figure 10.8. The increase in the number of quantitative educational research outputs

Evolution of number of quantitative outputs (based on articles with abstracts), 1995-2020

Note: This figure is presented to illustrate the shape of the trend rather than provide numbers. The calculations are based on the 56% of educational research outputs with an abstract
Source: Authors' calculations adapted from the LENS analytics set

StatLink 🔗 https://stat.link/hsm7yb

**Figure 10.9. The increase in the share of quantitative educational research**
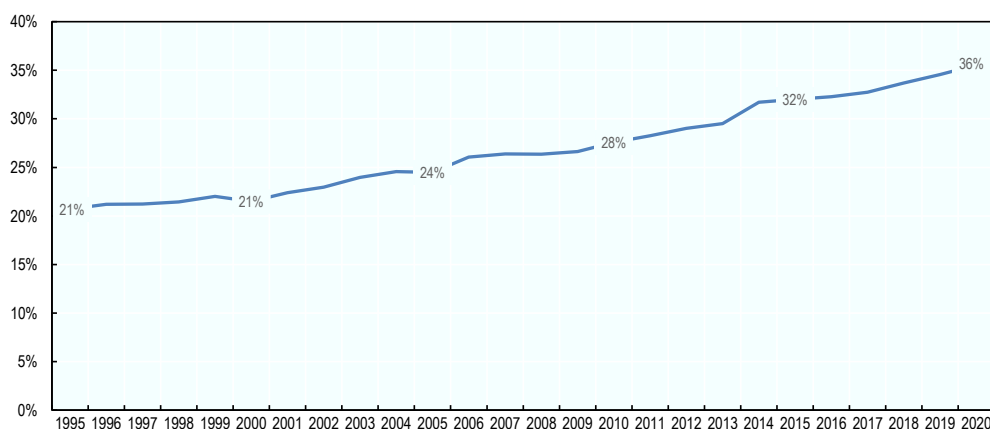
Source: Authors' calculations adapted from the LENS analytics set

An increasingly collaborative educational research

Finally, educational research is becoming increasingly collaborative, with an increasing number of research output jointly credited to several authors (Figure 10.10). While educational research was still less collaborative in terms of published outputs than other areas, it is catching up and went up from 42% to 68% between 1995 and 2020. By contrast, the share of collaborative output in all other domains has increased from 67% to 76% over the same period. We could unfortunately not compute the share of international collaborative research, which has increased as a general trend (OECD, 2017[1]). Our previous study based on SCOPUS (and thus having a slightly different methodology) showed a similar trend towards more collaborative output, including international collaborative output (Vincent-Lancrin and Jacotin, 2018[12]).

Figure 10.10. Share of educational research documents written in collaboration, 1995-2020
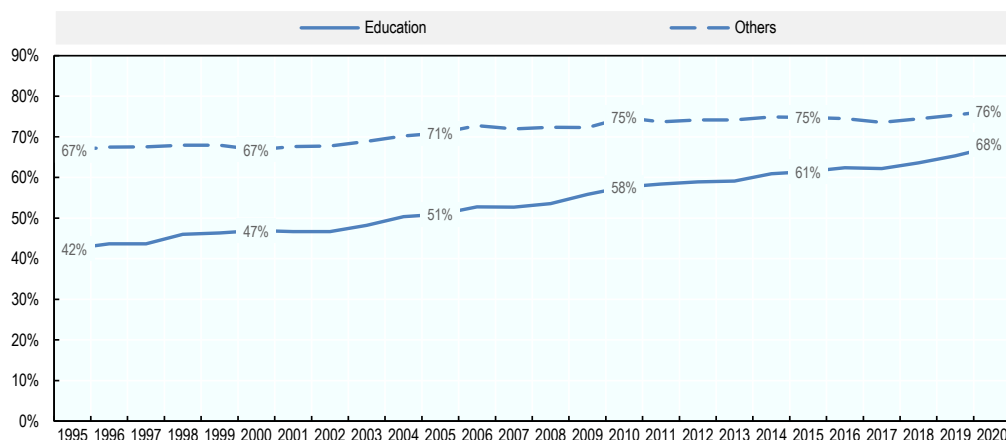
Source: Authors' calculations adapted from the LENS analytics set

Conclusion

The main purpose of this study was to show that valuable information about the educational research internationally can be derived from a bibliometric approach. Using a public bibliographic tool (LENS) with a very large coverage of academic publication, we found the following results:

- Educational research has increased both in quantity and as a share of the general research output;

- OECD countries produce the majority of the educational research output, but their share in the world output has decreased over the past decades;

- The United States is by far the first producer of educational research, but its share has also decreased over the past decades, with countries such as Brazil, Indonesia and China having significantly increased their output from a low starting point in the past few years;

- Educational research is mainly produced by researchers in the social sciences and in humanities, but half of it was produced in other fields of science in 2020, notably the health and natural sciences;

- Educational research has become increasingly quantitative in nature in the past decades, even though its qualitative output remains largely prevalent;

- Educational research is converging towards the same patterns as other disciplines in terms of collaboration (as measured by co-authored documents).

There are many other types of analyses that bibliometric approaches could make possible. While it was beyond the scope of this study, one could analyse the citation impact of different countries, the levels of inter-citation (and thus inter-connection) between different subfields of science, international collaboration, the extent to which patents cite educational research articles (and vice versa) and thus, how educational research connects with educational development. Even the "topics" researched might be captured through bibliometrics.

There are some limitations with bibliometric studies. The large amount of data and the incompleteness of the publication databases means that noise and measurement errors will tend to be large. It is thus recommended to run similar analyses with different databases before making strong conclusions. Work with both SCOPUS and LENS (using OpenAlex and other publication databases) shows that the results are relatively stable and that the general trends can be found independently of the database used (see Annex 10.A.). Using bibliometric data to compare the "performance" or "productivity" of different countries (or domains) may also be difficult.

First, there are language issues. While an increasing number of educational researchers publish in English, they still tend to publish in their country's language as their research may then have more impact on their education system. The multiplicity of languages in international comparisons complicates bibliometric approaches as ideally the research terms should be provided in all languages.

Second, the quantity of output between different fields of study may be misleading as different disciplines have different publication traditions (or expectations): in some disciplines it is common to publish a large number of research papers a year, whereas it is not the case in others. This can for example depend on the number of co-authors (as extensive co-authorship can lead to greater citation impact (Parish, Boyack and Ioannidis, 2018[13]). The standards are particularly different between health or computer science and social science (as social science research outputs typically involve less co-authors and are longer).

Third, it is very difficult to assess the quality of educational research through this means, even though this is what really matters at the end. The use of citation impact has its limitations as networks of low quality research may reach high citation impact. Limiting bibliometric analysis to a specific set of pre-defined "high

quality" journals is not fully convincing either as this does not allow for other poles of quality to emerge (nor reflect what may actually influence decision makers). Citation impact is also related to the nationality of authors or co-authors and thus give an advantage to large communities of researchers in a specific language, at least in fields such as educational research where not all the research output is published in English. For these reasons, bibliometric information should mainly be used to document the quantity rather than quality of the research output (or productivity and performance of countries).

Given the limitations mentioned above, trend data provide reasonably good comparative information. Without much investment, countries could produce yearly indicators about educational research within the OECD area and its partner economies, as is the case in other fields of science. Such information would for example allow countries with small educational research output to promote more production of educational research, help identify ways to bring together disconnected research communities and measure progress, identify collaboration opportunities, and evaluate the extent to which the quantity of research correlates with its use or with the quality of countries' education.

One of the methodological novelties of this chapter is to identify the educational research output through a semantic search. The approach adopted in this chapter can certainly be improved and such approaches have to be tailored to each database (and tool), but as education is both a "sub-field of science" and a subject of research for different fields (or "socio-economic objective"), we recommend to use this type of methodology. Only in this way can one capture the educational research output produced in fields of science that may have a different scientific tradition but contribute to understanding educational processes from different perspectives.

References

Brunner, J. and F. Salazar (2009), "La investigación educacional en Chile: Una aproximación bibliométrica no convencional", *Documento de Trabajo CPCE Nº 1*. [3]

Cretu, D. and Y. Ho (2023), "The Impact of COVID-19 on Educational Research: A Bibliometric Analysis", *Sustainability*, Vol. 15/6, https://doi.org/10.3390/su15065219. [7]

Diem, A. and S. Wolter (2013), "The use of bibliometrics to measure research performance in education sciences", *Research in Higher Education*, Vol. 54/1, pp. 86-114, https://doi.org/10.1007/s11162-012-9264-5. [6]

Gunnes, H. and K. Rørstad (2015), *Educational R&D in Norway 2013: Resources and Results*, NIFU, http://hdl.handle.net/11250/2412424. [4]

OECD (2022), *Who Cares about Using Education Research in Policy and Practice?: Strengthening Research Engagement*, Educational Research and Innovation, OECD Publishing, Paris, https://doi.org/10.1787/d7ff793d-en. [10]

OECD (2017), *OECD Science, Technology and Industry Scoreboard 2017: The digital transformation*, OECD Publishing, Paris, https://doi.org/10.1787/9789264268821-en. [1]

OECD (2007), *Evidence in Education: Linking Research and Policy*, OECD Publishing, https://doi.org/10.1787/9789264033672-en. [9]

Parish, A., K. Boyack and J. Ioannidis (2018), "Dynamics of co-authorship and productivity across different fields of scientific research", *PLOS ONE*, Vol. 13/1, https://doi.org/10.1371/journal.pone.0189742. [13]

Phelan, T., D. Anderson and P. Bourke (2000), *Educational research in Australia: a bibliometric analysis*, Higher Education Division, Department of Education, Training and Youth Affairs. [2]

Schneider, B., M. Carnoy, J. Kilpatrick, W. Schmidt and R. Shavelson (2007), *Estimating Casual Effects Using Experimental and Observational Designs*, American Educational Research Association, https://www.aera.net/portals/38/docs/causal%20effects.pdf. [11]

Sezgin, A., K. Orbay and M. Orbay (2022), "Educational Research Review From Diverse Perspectives: A Bibliometric Analysis of Web of Science (2011–2020)", *SAGE Open*, Vol. 12/4, https://doi.org/10.1177/21582440221141628. [5]

Vincent-Lancrin, S. (2023), *Public budget and expenditures in educational R&D: towards a new generation of international indicators?*, OECD Publishing. [8]

Vincent-Lancrin, S. and G. Jacotin (2018), *Education Research Indicators Worldwide*. [12]

Annex 10.A. Methodology

This annex presents the methodology to identify the educational research sample in more details and compares it with another approach that was tested by the authors in a previous study. Vincent-Lancrin and Jacotin (2018[12]) developed a methodology on the SCOPUS bibliometric database to identify the educational research output through a semantic search. The method was adapted to the LENS database and tool. The first section of this annex presents the initial methodology, then shows how it was adapted to LENS, before comparing the results of the two approaches on the period that is common to the two studies.

Past methodology with SCOPUS

The methodology that was used with the SCOPUS database allowed for a nuanced semantic search and categorisation of the educational research output. At the time of the study, in 2016, SCOPUS covered more than 21 500 peer-reviewed journals across the world and provided information about research output such as title, abstract, author's affiliations, year of publication, type of publication, discipline (SCOPUS classify journals in broad fields of science: social sciences, health sciences, physical sciences and life sciences), etc. The considered only articles published since 1996 (more than 60% of the full database). (The OECD and UNESCO – United Nations Educational, Scientific and Cultural Organization – databases on education enrolment and R&D expenditures' in higher education were also used to compute different indexes.)

In order to determine which research articles could be considered as educational search, we carried out a semantic search in the title and abstract fields of the research articles. Through an iterative process, we defined a set of search terms related to education, separated in two different classes, which could be "strongly" or "weakly" related to education topics. The list of specific terms is the following:

- "Strong" terms: *educat* (in English, French, Spanish & *bildung* for German) without prefix, *student*, *teach*, *school* (*schul* in German) without prefix and suffix, *academ* (*akadem* in German), *curricul*, *classroom*, *pedagog* (*padagog* in German), *campus* without prefix, and *kindergarten*;
- "Weak" terms: *learn* (*lehrer* in German), *gradu*, *traini* or *traine* without prefix, *instruct*, *college*, *facult*, *cognit* without prefix, *intel*, *tutor* and *didact*

Annex Table 10.A.1. Number of identified education articles by search criteria

	Title		Abstract		
	Strong	Weak	Strong	Weak	Total
At least	1	0	0	0	474 836
At least	0	1	1	2	65 300
At least	0	0	3	0	53 522
At least	0	0	2	1	50 279
Total					643 937

Source: Authors' calculations adapted from the SCOPUS database

To identify education articles, we drew different random samples of articles from more restrictive to less restrictive conditions and iteratively constructed the semantic search that yielded the most satisfactory results. Through this process, we retained as educational research articles those that included:

- At least one "strong" term in the title; or
- At least one "weak" term in the title, one "strong" term and two "weak" terms in the abstract; or
- At least three "strong" terms in the abstract; or
- At least two "strong" terms and one "weak" term in the abstract.

The number of articles retained by applying this search strategy is reported in Annex Table 10.A.1.

As shown in Annex Table 10.A.2, the terms "*educat*" and "*student*" are those with the highest occurrence within the 643 937 articles of education.

Annex Table 10.A.2. Occurrences of each term among articles in education

	Title	Abstract	Total	%
Strong words				
educat	172 675	300 043	357 074	55.5%
student	122 869	304 615	325 913	50.6%
teach	102 996	207 029	238 099	37.0%
school	55 928	124 199	144 460	22.4%
academ	43 514	98 540	118 044	18.3%
curricul	22 219	74 831	81 893	12.7%
classroom	17 797	55 055	61 563	9.6%
pedagog	10 067	36 643	40 394	6.3%
campus	5 108	13 063	14 779	2.3%
kindergarten	2 328	5 363	5 836	0.9%
Weak words				
learn	77 308	191 570	202 608	31.5%
gradu	22 640	82 851	88 978	13.8%
train(i/e)	22 667	81 025	85 134	13.2%
instruct	11 520	59 695	61 813	9.6%
college	18 849	53 709	57 772	9.0%
facult	6 920	41 083	42 346	6.6%
cognit	7 394	41 092	41 763	6.5%
intell	6 903	21 528	22 465	3.5%
tutor	3 768	12 801	13 158	2.0%
didact	793	6 962	7 179	1.1%

Source: Authors' calculations adapted from the SCOPUS database

Some search terms initially in the list above were later on removed. This is the case of the term "*university*" which, despite its clear connection to higher education, is often used at the end of the authors' title or in the abstract for the purpose of listing the institutional affiliation of the authors. This term was present in the title or abstract of 790 694 articles. Only 7% of them (55 418) were published in journals of education and among them, 42 047 (76%) were included in our education sample according to the criteria defined above. In comparison, the term "*educat*" (702 412) is less frequently used but 21% of articles containing this term are published in journals of education. This is three fold more frequent than the term "*university*".

In full count, there are 386 662 articles published in journals of education. 209 861 (54%) among them are included in our education sample. However, it is important to underline that around 75 000 articles in journals of education don't have an abstract (used as criteria of inclusion, except for the most restrictive criteria), reducing their probability of inclusion. Among them, only 30% are included thanks to their title.

Methodology with LENS

The methodology used to identify the sample of educational research output is presented in the first section of the paper. It very much followed the idea of the previous semantic search, but in a simplified way.

Annex Table 10.A.1 presents some details about the LENS database, notably the number of papers with an affiliation (this a country) and the characteristics of the analytics set that is designed for bibliometric studies and includes papers with more affiliations and abstracts. While the database includes a variety of scholarly documents, the large majority are journal articles and book chapters.

Annex Table 10.A.3. Characteristics of the scholar work in the LENS database

	Number	Percent
Scholar work in the database	**236 413 556**	
With affiliation	80 364 549	34%
Scholar work in the analytics set of the database	**107 987 328**	
With affiliation	64 182 466	59%
With abstract	66 513 909	62%
Type of publication in the analytics set of the database	**107 987 328**	
Journal article	88 705 081	82%
Book chapter	10 727 488	10%
Conference proceedings	4 326 786	4%
Conference proceedings article	2 509 192	2%
Book	1 604 529	1%
Letter	94 280	0%
Review	19 972	0%
Type of information in the education analytics set of the database	**2 629 809**	
With affiliation	1 376 458	52%
With abstract	1 478 685	56%

Source: LENS

Annex Table 10.A.2 presents the different sequences of characters of works that were used to identify the educational research corpus. In the case of LENS, just working on the titles rather than both title and abstract was much easier and yielded similar results to the more complex criteria used on the SCOPUS database. The method allowed us to identify 2.6 million educational research papers. Random sub-samples were checked manually to assure that the output qualified as educational research in the sense that education was one of its subjects.

Annex Table 10.A.3 shows the sequence of characters that have been used for the search to identify quantitative papers using the documents' abstracts. The analysis with SCOPUS also included some characters that could not be used in LENS ("=", "<", ">" and "%"). The same methodology that had been validated was then reused and the selection included documents with either a strong word in its abstract

(432 091) or two weak ones (296), leading to identifying a total of 432 377 documents with a quantitative methodology (which sub-samples checked manually).

Annex Table 10.A.4. Identification of the education analytic set

	Character sequence	Title	%
	educat	903 524	34.4%
	educac	98 877	3.8%
	bildung	18 661	0.7%
Educat		**1 015 269**	**38.6%**
	student	579 567	22.0%
	teach	527 291	20.1%
	school	494 086	18.8%
	schul	295	0.0%
School		**494 378**	**18.8%**
	academ	210 312	8.0%
	akadem	17 463	0.7%
Academ		**227 329**	**8.6%**
	curricul	99 679	3.8%
	classroom	81 451	3.1%
	pedagog	83 445	3.2%
	padagog	6 908	0.3%
Pedagog		**90 326**	**3.4%**
	campus	25 584	1.0%
	kindergarten	10 811	0.4%
Total		**2 629 809**	

Source: Authors' calculations adapted from the LENS analytics set (30 August 2021)

Annex Table 10.A.5. Identification of the quantitative educational research output

	Character sequence	Abstract	% education research	% quantitative
strong words				
	data	303 276	20.5%	70.1%
	sample	133 311	9.0%	30.8%
	statistic	85 493	5.8%	19.8%
	control group	50 144	3.4%	11.6%
weak words				
	quantitativ	46 989	3.2%	10.9%
	estimat	33 392	2.3%	7.7%
	gauss	75	0.0%	0.0%

Source: Authors' calculations adapted from the LENS analytics set (30 August 2021)

Robustness check: comparing the results of the two studies

While one of the well-known weakness of bibliometric studies lies in the strengths and weaknesses of each database, the use of two different ones allowed us to compare results. They show that while the levels can be slightly different for the two studies, the trend is similar.
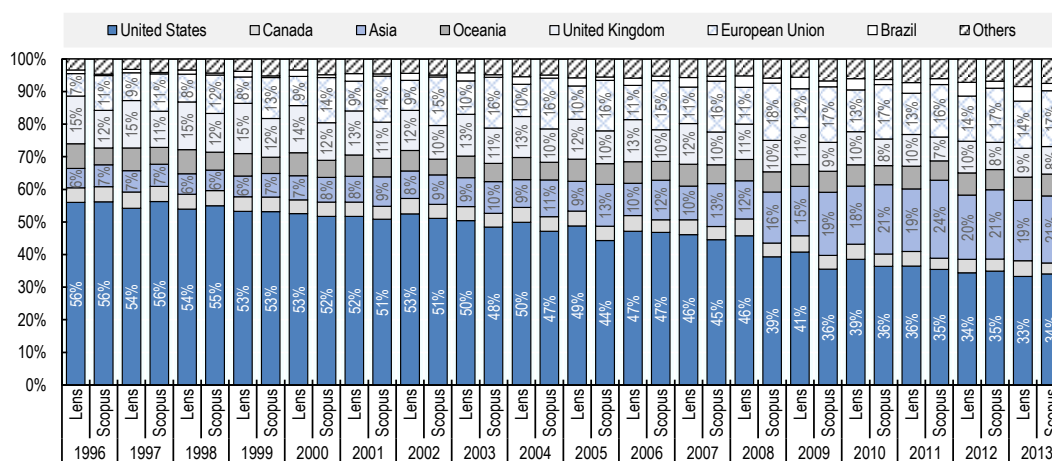
Annex Figure 10.A.1. Evolution of the share of the educational research output in the total research output based on the two studies
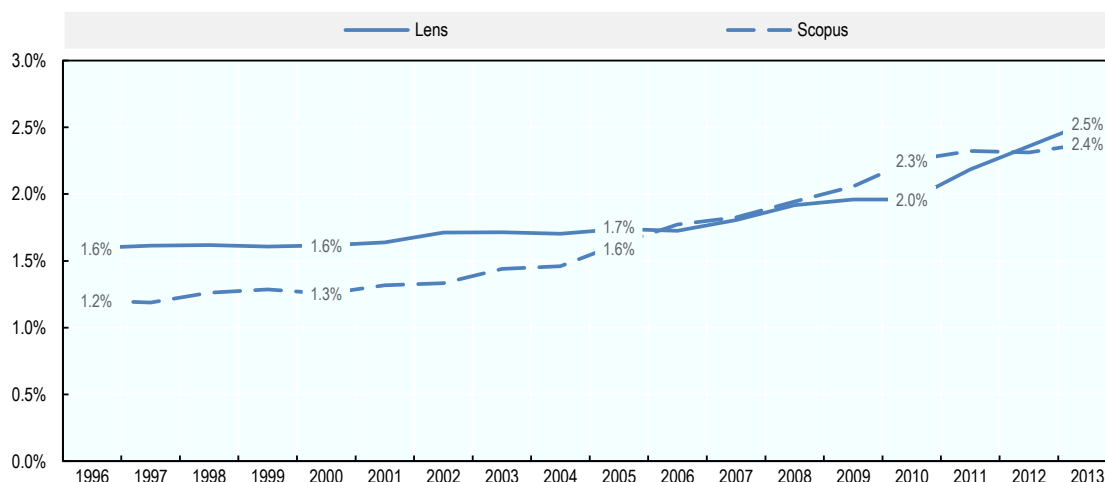
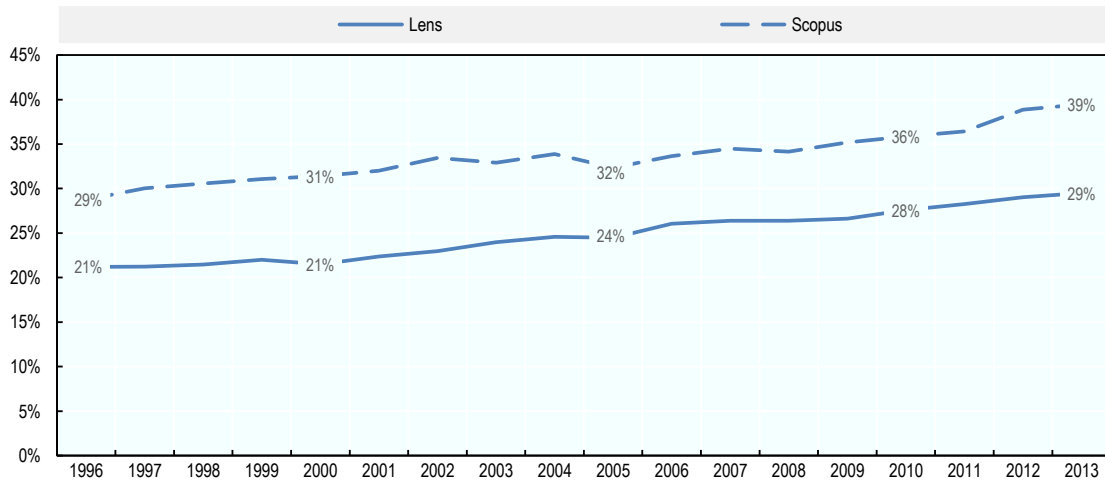Source: Authors' calculations adapted from the LENS and SCOPUS databases

Annex Figure 10.A.2. Evolution of the share of countries/regions' output in the world educational research output based on the two studies

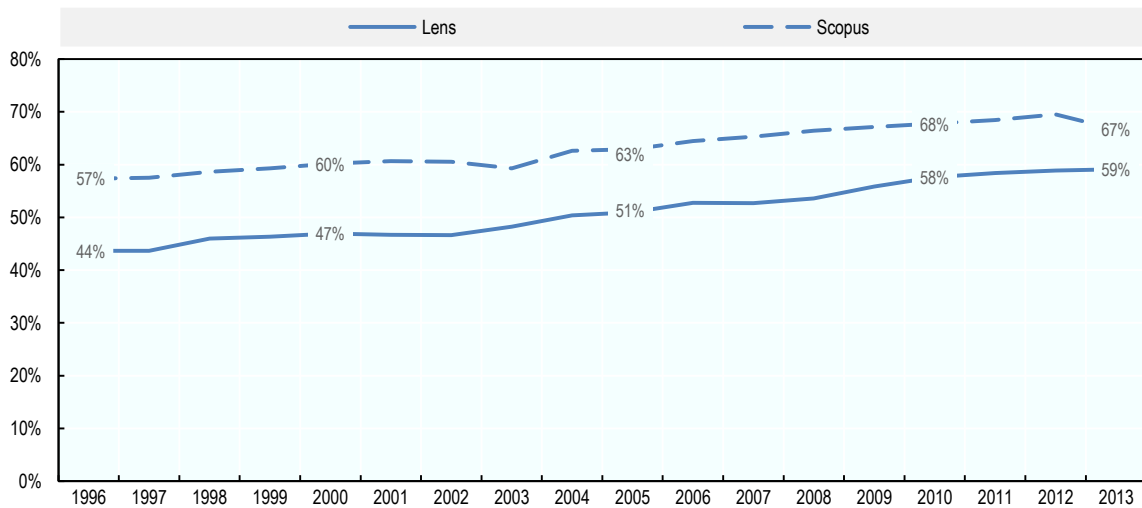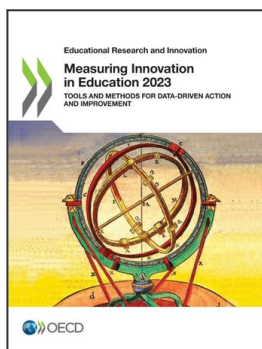Source: Authors' calculations adapted from the LENS and SCOPUS databases

Annex Figure 10.A.3. Evolution of the share of quantitative educational research output based on the two studies

Source: Authors' calculations adapted from the LENS and SCOPUS databases

Annex Figure 10.A.4. Evolution of the share of collaborative educational research output based on the two studies

Source: Authors' calculations adapted from the LENS and SCOPUS databases

Notes

[1] See https://www.oecd.org/sti/scoreboard.htm#explore for the OECD STI Scoreboard platform.

[2] For time and budget reasons, we simplified the methodology that we developed and piloted for a previous, unpublished study using another bibliometric database: the educational research output was effectively identified through a conditional method that is also presented in Annex 10.A and included a mix of "strong" and "weak" words (or strings of characters) in the title and/or abstract of the papers for the search. In the case of LENS, the addition of conditions did not seem to make much difference to the research output (but increased the time and complexity of the queries) which led to a simplification of the methodology.

[3] The presented classification here is based on the university affiliation of the authors. The public LENS application does not allow one to do fractional counts. However, a comparison with a previous unpublished research based on SCOPUS implementing fractional counts shows that the results are similar.