Discussion Paper Series

RIEB

Kobe University

DP2019-27

A Review of Data Used in Education Research: Focus on Empirical Studies in Developing Countries

Phal CHEA

December 26, 2019

Research Institute for Economics and Business Administration

Kobe University

2-1 Rokkodai, Nada, Kobe 657-8501 JAPAN

A Review of Data Used in Education Research: Focus on Empirical Studies in Developing Countries

Phal Chea*

December 2019

Abstract

This paper intends to review education-related datasets, including data from household surveys, learning assessments and field experiments, publicly available for researchers and students interested in conducting education research. It also presents ideas on how those data can be used in empirical studies and identifies some major potential sources of those datasets. Issues in education have shifted from access to quality learning and at the same time, randomized control trial (RCT) has become the gold standard in measuring the impacts of education programs. The paper also notices the emerging field of educational data mining employed to predict student performance or to identify at-risk students.

Keywords: Education data, household survey, learning assessment, field experiment, educational data mining

JEL Codes: A20, C80, I21, I25

* Adjunct Researcher, Research Institute for Economics and Business Administration (RIEB), Kobe University

1. Introduction

Despite good progress in expanding educational access over the past decades in developing countries around the world, it is estimated that more than 120 million children and 137 million adolescents and youth are still out of school in 2018 globally (UIS, 2019). The progress in reducing the global number of out-of-school children has stagnated in recent years, regardless of the adoption of Education for All (EFA) in 2000 and Sustainable Development Goal 4 (SDG 4) in 2015 aiming at realizing universal primary and secondary education. In addition, recent studies indicate that many children fail to acquire basic knowledge and skills needed to be productive after leaving school. As a result, there has been a shift from the emphasis on access (years of schooling) toward the quality of learning using learning achievement as a benchmark. At the same time, there is an increasing number of studies in the past two decades employing more rigorous research designs using the field experiment approach to measure casual impacts rather than relationships between education interventions and education outcomes. Educational Data Mining is another new emerging interdisciplinary research field, combining computer science, statistics, and education, to better understand and predict students' success or failure.

To monitor progress toward SDG 4 and to make well-informed decisions based on evidence, reliable and high-quality data are indispensable. Key sources of education data include administrative datasets collected through education system, household surveys, learning assessments, and data collected through specific education projects. This paper intends to review education-related datasets which are publicly accessible for researchers and students interested in conducting research in the field of education development. It also provides some examples of existing literature that have employed those datasets in the empirical analysis as well as identifying the potential sources where those datasets can be obtained.

2. Household Surveys and Population Censuses

Household surveys are an important source of data on educational access, participation, and attainment. Most countries conduct household surveys to collect a wide range of information from health, education, employment to consumption on a regular basis. In most cases, information such as education level, grade completion, school participation and literacy of household members aged five and older can be found in household surveys (Education Policy and Data Center, 2009). Some household surveys, including the Living Standards Measurement Study (LSMS), also provide information about household expenditure on education. Combining with other information of

households characteristics (e.g. household size, gender of household head, parental education, and household wealth) and of individuals (e.g. age, gender, relationship to household head, and employment status), household surveys are frequently used to assess the relationships between education access and other demand-side factors to understand the education system as well as to identify educational issues of particular populations. However, since the scope of household surveys is very broad, normally information related to supply-side factors, such as characteristics of schools and teachers, is scarcely included. Supply-side information can be obtained from education administrative statistics collected by local governments from schools, such as Education Monitoring Information System (EMIS), but in most cases, it is extremely difficult to merge household data and school administrative data for analysis. In addition to household surveys, national statistics offices also conduct national censuses, albeit less frequently conducted, and the number of variables found in the national census is much smaller.

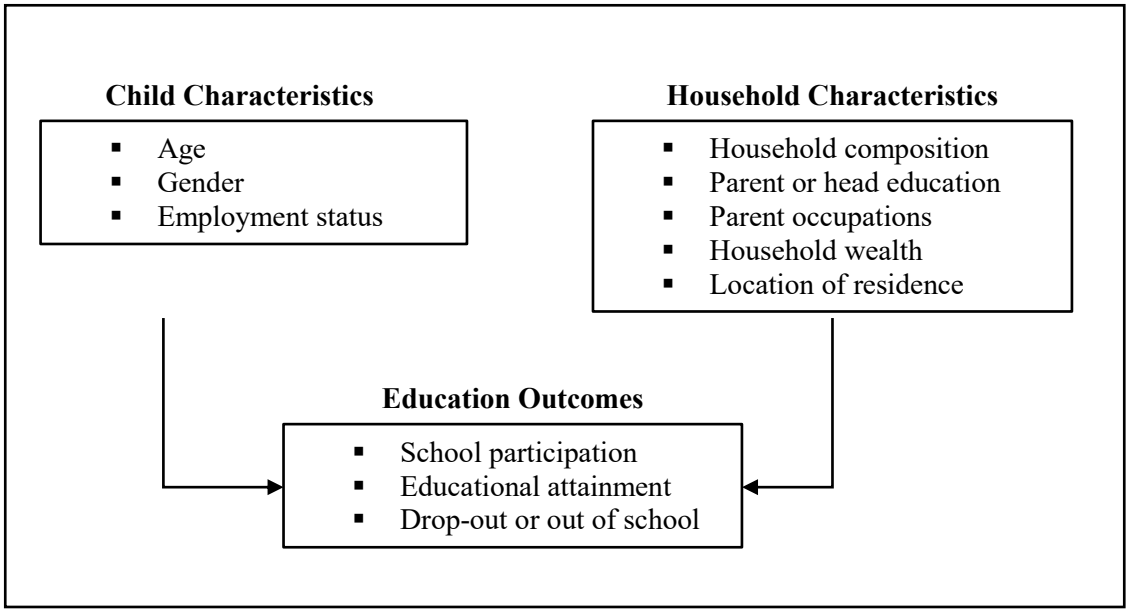
2.2. Studies Using Household Survey Data in Education

Household surveys or national censuses are a good source for the estimation of access to education, numbers of out-of-school children, and/or dropout students. There is also a bulk of literature of empirical studies using household surveys to examine how demand-side factors influence children's school participation, education attainment or dropouts in developing countries as illustrated in Figure 1. These kinds of analyses can help identify individual and household factors associated with school participation and education attainment or shed some light on educational gaps between subgroups of populations. Household factors commonly found to have influences on children's schooling include parental education, socioeconomic status, and residence location (Connelly & Zheng, 2003; Mani, Hoddinott, & Strauss, 2013; Sánchez & Singh, 2018; Tansel, 2002). Studies on the effect of migration on the schooling outcomes of left-behind children have produced mixed results as migration is inextricably linked with remittances that can off-set the absence of migrant household members (Iwasawa, Inada, & Fukui, 2014; Mansuri, 2006; Rapoport & Docquier, 2006). Although the correlations found between labor status and schooling outcomes are rather controversial, long-hour child labor or child work seems to do more harm than good to the human capital accumulation of the children (Edmonds, 2007; Guarcello, Lyon, & Valdivia, 2015).

Since household surveys normally also collect information of household income and expenditure, we can also conduct a benefit incidence analysis to examine the distribution of public education spending across the population to assess inequality issues by assuming that education expenditure per student of each level is the same (Lassibille & Tan, 2007). For instance, using Ghana

Living Standards Survey (GLSS5) 2005/06, Gaddah, Munro, and Quartey (2015) found that the poor in Ghana greatly benefit from the increase in public spending, in particular in the pre-school and primary school levels. The Mincer equation, using wage earnings as a function of schooling and working experience, is widely adopted in the estimation of the rates of return to education (Psacharopoulos, 1994; Psacharopoulos & Patrinos, 2018). The cross-country study in 1994 revealed that return to primary education was very high, but the study using newer datasets in 2018 found that return to higher education has become higher recently. Information about hour wages, education backgrounds, and working experiences are commonly available in household surveys or labor surveys for such a rate of return analysis.

Figure 1. Analysis Framework of Household Factors and Educational Outcomes



Source: Created by the author based on UIS (2004)

Although information on government's educational interventions is scarce in household surveys, some interventions, such as government's scholarships and school feeding programs, can be found in some household surveys. Canton and Blom (2004) examined the relationship between a student loan program and university enrollment using a probit model with Mexico's national household survey data. While it is easy to investigate relationships between public interventions and educational outcomes using cross-sectional household surveys, it is much more challenging to measure the impact (causality) of the programs due to self-selection issues. In other words, characteristics of public program beneficiaries are normally different from non-participants. Randomized Controlled Trial (RCT), to be discussed later, is now the gold standard to evaluate the

program impacts. Nevertheless, there are some techniques using econometric models, such as instrument variables (IV) and propensity score matching (PSM), to capture the impact of such program interventions (Angrist & Pischke, 2008). For example, Nakata (2013) uses the IV method with data from the Vietnamese Household Living Standard Survey (VHLSS), in an attempt to measure the effect of government's financial assistance programs on access to higher education among upper secondary graduates from poor households.

2.2. Potential Data Sources

Some well-known large-scale international household surveys include the Demographic and Health Survey (DHS), the Multiple Indicator Cluster Survey (MICS), and the Living Standards Measurement Study (LSMS). Countries conducting these household surveys are listed in Table A.1 in the Appendix.

- **The Demographic and Health Survey (DHS)**: Started in the 1980s, DHS has a big collection of more than 400 household surveys collected from over 90 countries, providing good sources of information about health, nutrition, demography, socioeconomic status, and education. Due to its large sample size, in the range of 5,000 to 30,000 households, it can be represented by both the national and sub-national levels. Recent surveys also provided the GPS data of the sampled locations. Different from other surveys, DHS has detailed information on women aged 15-49 years old. DHS is supported by ICF International and the U.S. Agency for International Development (USAID).[1]

- **The Multiple Indicator Cluster Survey (MICS)**: Although MISCS was only started in 1995, it has expanded rapidly in a short period of fewer than three decades. As of 2019, UNICEF has conducted 326 surveys in more than 116 countries around the world, mainly in developing countries. It has become an important source of internationally comparable data to produce indicators on women and children to track the agreed targets of the Millennium Development Goals (MDGs) and Sustainable Development Goals (SDGs). Variables related to topics such as maternal and child health, education, child mortality, and child protection can be found in the MICS survey.[2]

[1] https://www.dhsprogram.com/data/
[2] https://mics.unicef.org/surveys

- **The Living Standards Measurement Study (LSMS)**: Initiated in 1980 aiming to improve the quality of household data, the World Bank has worked in collaboration with local government statistical offices in developing countries to collect household data and, at the same time, built up the capacities of the national statistical officials. Key indicators include consumption, income, savings, human capital, and anthropometrics. To date, 37 countries have participated in the LSMS, and 121surveys are listed in the LSMS collection of the World Bank's microdata library.[3] The sample size of the LSMS is relatively small, around 2,000 to 5,000 households, as the World Bank puts more effort into the quality of the data. Although it is nationally representative, in most cases, it is not designed to represent the sub-national level. In the LSMS surveys, households are asked to report their expenditure on the education of their members enrolled in formal education systems from pre-primary school to post-secondary school; however, it does not cover technical and vocational education training (TVET) and non-formal education (UIS, 2017).

Besides the above-mentioned large-scale household surveys, there are smaller-scale household surveys, including Integrated Household Survey (IHS), Core Welfare Indicators Questionnaire surveys (CWIQ), Household Income and Expenditure Survey (HIES), and Rand Family Life Surveys (FLS) as well as country-specific national household surveys. However, these survey data sets, in particular, the country-specific household surveys, are rather difficult to obtain. The International Household Survey Network (IHSN) has collections of 6,986 surveys as of the end of 2019 from most countries in the world.[4] Although the data sets are not available directly from the website, information on how and from whom those data sets are shared on their online database. IPUMS-International houses the largest collection of publicly available census samples and registered users can download sub-sets of the samples of those national censuses for free.[5] However, the number of variables is much less abundant in comparison to the household survey.

3. Student Learning Assessments

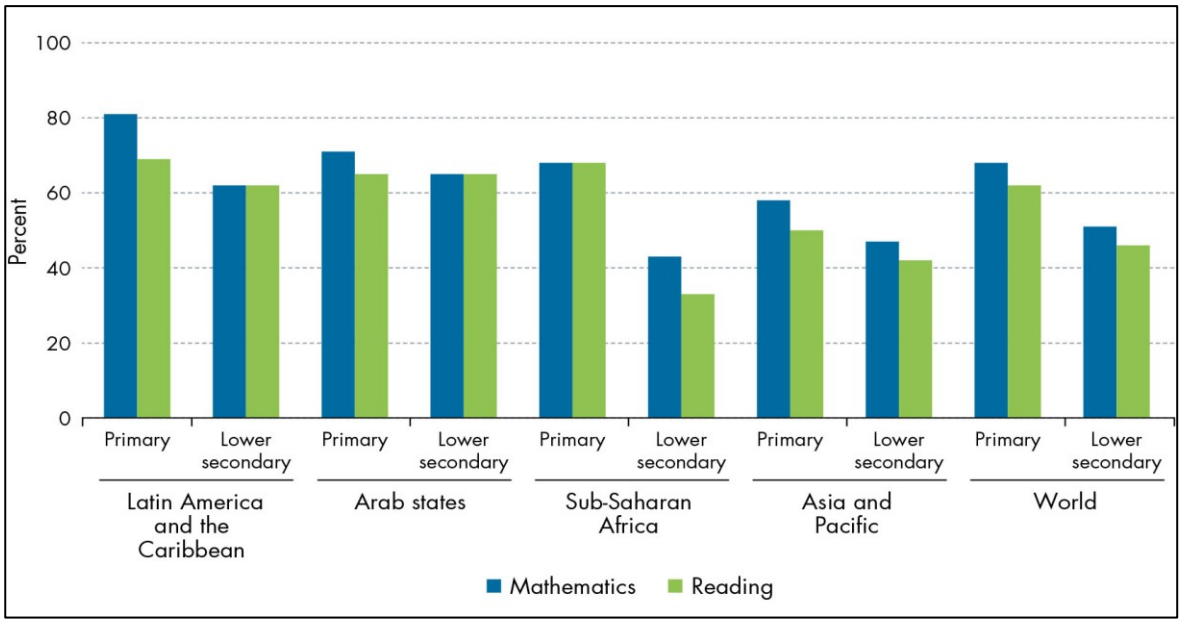
Educational attainment used to be measured by years of schooling or grade completion, and educational intervention traditionally emphasized expanding access to education, but a growing number of studies now suggest that it is not the number of years students stay in school, but what they actually learn in that period that matter the most (Hanushek & Woessmann, 2011; Pritchett,

[3] https://microdata.worldbank.org/index.php/catalog/lsmst
[4] http://catalog.ihsn.org/index.php/catalog
[5] https://international.ipums.org/international/

2013). UNESCO (2013) alerted, in its annual Global Monitoring Report, that a large number of children and young people leave the education system without the competencies they need to lead productive and healthy lives. The World Bank (2018) dedicated three chapters of its annual flagship report, the World Development Report 2018, talking about the learning issue. In response to this learning crisis, effective learning outcomes are now included as indicators in SDG 4 to ensure learning quality for all. Learning assessments can be designed and implemented internally within classrooms or externally through standardized national, regional or international assessments. However, this paper only covers the later standardized learning assessments.

There are critiques that overemphasize assessment, normally in developed countries, which fails to prioritize actual learning and that, in some cases, teachers and schools even engage in cheating to increase their students' performance (Fausset, 2014; Jacob, 2005). In contrast, in many developing countries, there is no reliable information to measure and monitor whether program interventions or polices actually translate into learning (World Bank, 2018). Although UNESCO (2015) reports that the number of countries conducting national learning assessments increased from 12 in 1990 to 101 in 2013, the percentage of countries with reliable data to track their progress toward SDG 4 regarding learning is still very low (see Figure 2). Less than half of the countries in Asia and the Pacific region have reliable data on learning outcome indicators to monitor their progress toward SDG 4, meaning that there are many countries still have no means to measure or evaluate student performance and effectiveness of national education programs.

Figure 2. Percentage of Countries with Reliable Learning Assessment Data at Primary and Secondary Schools



Source: World Bank (2018): 17

Well-designed learning assessments can be used to inform policymakers of the disparities among gender, socio-economic, region and other characteristics and to reveal education system performance trends. Besides student performance (test scores), learning assessments also include additional questionnaires to gather information related to the characteristics of students, families, teachers, schools, and communities. Most of the large-scale international assessments are designed to offer a wide range of variables such as student learning environment at home and school, family background, socio-economic status, and school resources (Cresswell, Schwantner & Waters, 2015). In addition to the school-based learning assessments, there are household-based assessments conducted by Non-Government Organizations (NGOs), such as the Annual Status of Education Reports (ASER) in India and Pakistan, and UWEZO in Kenya, Tanzania, and Uganda. Different from school-based assessments, these assessments are able to include out-of-school children in their samples by visiting households directly.

3.1. Studies Using Learning Assessment Data

The large-scale learning assessments at regional and international levels are designed for various purposes, but oftentimes they are to evaluate the quality of the education system, to ensure equity among sub-groups of students, and to identify possible interventions to improve the learning outcomes and equity (Tobin et al., 2015). International assessments, using the same standardized tests across participating countries, allow researchers to make comparisons between the countries and track the progress of the countries that participated in the same assessment on several occasions (Crawfurd et al., 2019). Studies using learning assessment data can help educators and policymakers make more informed decisions and guide them in formulating new policies or improving the existing ones (Raudonyte, 2019).

Numerous studies have tried to identify the factors–family background, home input, teacher quality or school resources–that explain the gaps in student performance (Ammermüller et al., 2005; Fuchs & Wößmann, 2008; Glewwe et al., 2011; Hanushek & Woessmann, 2011). For instance, Ammermüller et al. (2005) examined the TIMSS data of seven countries in Eastern Europe and concluded that student backgrounds (such as gender, age, and immigrant backgrounds) have stronger effects on students' learning achievement than school resources do. Nevertheless, a systematic review conducted by Glewwe et al. (2011) on the effects of school resources on student's learning in developing countries indicated that school infrastructure, for instance, good quality of roofs and

walls, tables and chairs, school libraries, teachers' subject knowledge, and teacher in-service training, have positive effects on students' learning outcomes.[6]

A large share of education's recurrent budget is spent on teacher salaries, meaning that reducing or increasing class size (number of students per class) significantly influences the education budget. Yet, it seems that small class size does not always result in good learning outcomes. Based on TIMSS's learning assessment data of 11 countries, Wößmann and West (2006) found that in some countries small class size does not lead to better student performance in mathematics and science test scores. Studies using SEACMEQ II, the regional student assessment in Sub-Sahara Africa, revealed that streaming students by ability level lead to wider performance gaps. Based on the results, Seychelles decided to revise the national policy to terminate the student streaming at all education levels (Leste, 2005).

3.2. Potential Data Sources

International and regional learning assessment includes not only students' test scores but also information related to the characteristics of students, parents, teachers, and schools. As mentioned earlier, more and more countries are interested in measuring the students' learning outcomes. Below are some potential sources of learning assessment datasets that are publicly available for free. A list of other learning assessments can be found in Table A.2 in the Appendix.

- **Programme for International Student Assessment (PISA)**: Conducted every three years since 2000, PISA was developed by the OECD to assess the knowledge and skills of 15-year-old students in reading, mathematics, and science. The first assessment was joined by 43 countries, mostly OECD member countries and other developed countries. In 2018, more than 600,000 students from 79 countries participated in the seventh PISA including new countries from Belarus, Bosnia and Herzegovina, Brunei Darussalam, Morocco, the Philippines, Saudi Arabia, and Ukraine. Although PISA has expanded to include more countries from lower-middle income countries, none of the participants are from low-income countries (Crawfurd et al., 2019). There are also countries that previously joined the assessment but later dropped out. The PISA data sets are available for download from its website at https://www.oecd.org/pisa/data/. To increase the participation from middle-income and low-income countries, OECD launched the PISA for Development (PISA-D)

[6] 79 studies and papers were selected from 9,000 studies after several screening processes for the reviews. Not all, but a large proportion of the reviewed studies used assessments tests.

initiative in 2014. By 2017, eight countries (Bhutan, Ecuador, Honduras, Paraguay, Zambia, Cambodia, Guatemala, Panama, and Senegal) signed the participation agreement with the OECD and seven of them (except Bhutan and Panama) took part in the pilot assessments.

- **Trends in International Mathematics and Science Study (TIMSS) and Progress in International Reading Literacy Study (PIRLS):** Both TIMSS and PIRLS are directed by the International Association for the Evaluation of Educational Achievement (IEA) (Altinok, Angrist, & Patrinos, 2018). Started in 1995, TIMSS is held every four years to assess fourth- and eighth-grade students in mathematics and science. 580,000 students from 57 countries and 7 benchmarking entities participated in the sixth TMISS in 2015. Six years after TIMSS was introduced, PIRLS was conducted for the first time in 2001 to assess the reading ability of fourth-grade students. 61 countries and entities joined the fourth PIRLS in 2016. Very few countries from Latin America and Africa participated in these assessments. TIMSS and PIRLS datasets can be downloaded from https://timssandpirls.bc.edu/databases-landing.html.

- **Southern and Eastern Africa Consortium for Monitoring Educational Quality (SACMEQ):** In addition to international learning assessments, there are also regional assessments conducted in specific regions including SACMEQ in Southern and Eastern Africa, LLECE in Latin America, PILNA in Pacific Island, SEA-PLM in Southeast Asia, and PASEC in Francophone countries. Compared to international assessments, they are organized less regularly. Although most African countries do not participate in international students' learning assessments, many African countries join the regional assessment, SACMEQ. First conducted between the period of 1995 and 1999, so far three rounds of SECMEQ have been carried out to assess sixth-grade students in reading, mathematics, and sciences. The 15 participating countries of the third SACMEQ 2006-2011 are Botswana, Kenya, Lesotho, Mauritius, Malawi, Mozambique, Namibia, Seychelles, South Africa, Swaziland, Tanzania, Tanzania, Uganda, Zambia, and Zimbabwe. Registered users can make a request for the SECMEQ dataset through its website.[7]

[7] http://www.sacmeq.org/?q=seacmeq-data

4. Field Experiment Data (Randomized Controlled Trials)

The 2019 Nobel Prize in economics was awarded to Abhijit Banerjee, Esther Duflo and Michael Kremer for their innovative approach using field experiments, also known as Randomized Control Trial (RCT), to answer global development issues from agriculture, micro-finance to education. RCT is considered as the most rigorous approach in assessing the impacts of specific programs or interventions in development studies including in the field of education. It had been challenging to identify reliable causal impacts of a program due to the selection bias issue. However, this selection bias problem can be solved by randomly assigning individuals or groups of individuals into treatment and control groups (Duflo, Glennerster, & Kremer, 2007). The rise of evidence-based studies has helped governments identify interventions that can produce desired results or outcomes. The United Kingdom even created a research team, called the Nudge Unit, to conduct various field experiments to provide evidence-based results to the government to design new public policies (Thaler & Sunstein, 2008).

A range of educational interventions has been made to ensure that no one is left behind and that students in schools can acquire certain knowledge and skills. However, not all education programs are effective in producing the expected outcomes and some are just too expensive. Randomized evaluation has been used to identify which programs are effective so that governments can make better-informed decisions on education investments. As the field experiment approach has gained more popularity, there is an increasing number of evidence-based studies using it in education research. Connolly, Keenan, and Urbanska (2018) conducted a systematic review of education literature using the RCT method and found that more than 1,000 unique RCTs were implemented during the period of 1980 and 2016 around the world. More than two-thirds of those experiments happened in the last ten years after 2007. Interestingly, most of the field experiments in education are conducted in developing countries. International Initiative for Impact Evaluation (3ie), an International NGO promoting evidence-based development policies around the world, conducted a systematic review of studies on the impact of education interventions on access to education and learning outcomes in 52 lower- and middle-income countries and identified 238 experimental or quasi-experimental studies assessing 216 education interventions implemented between the period of 1990 and 2015 (Snilstveit et al., 2015).

4.1. Studies Using Experiment Data

The 3ie's systematic review categorized interventions into children, household, teacher, school and system levels and found that the most popular intervention is cash transfer (49 interventions) at the

household level (see Table 1). Normally, cash transfers are provided to a child's parent (mother is more preferable to father) with some conditions attached, i.e, the child needs to maintain a minimum attendance rate and/or not work and focus on learning. The review found that cash transfer has strong and positive impacts on school participation, but has little or no effect on student's learning (Snilstveit et al., 2015). The most common educational access outcomes include enrollment, attendance, and dropout, while the major learning outcome indicators are language, mathematics and art test scores. It seems that interventions aiming at addressing non-cognitive knowledge and skills are very scarce. Among the reviewed studies, a score of papers is related to pedagogical intervention at school level to improve students' learning outcomes by adapting or introducing new classroom practices, curriculum or training school teachers on how to use new teaching and learning materials. The meta-analysis suggested that most of the pedagogical interventions have positive effects on student test scores and help lower student dropout rates.

Table 1. List of Educational Interventions Reviewed by 3ie

Intervention Level	Intervention	No. of Studies
Child Level	School-Based Health	16
	School Feeding	16
	Merit-Based Scholarships	11
	Providing Information	4
Household Level	Reducing/Eliminating Fees	9
	Cash Transfer	49
Teacher Level	Teacher Hiring	8
	Teacher Incentives	10
	Teacher Training	1
	Diagnostic Feedback	2
School Level	Computer Assisted Learning	18
	Pedagogy	22
	Extra Time	3
	New Schools and Infrastructure	7
	Providing Materials	4
	Remedial Education	4
	Grade Retention	1
	Tracking	2
System Level	School-Based Management	14
	Community-Based Monitoring	11
	Public-Private Partnerships	13

Source: Snilstveit et al. (2015).

4.2. Potential Data Sources

Datasets of field experiment surveys, designed to evaluate specific programs or interventions, were not widely available for public scrutiny. However, an increasing number of researchers and research institutes have committed to research transparency and have started to provide unrestricted access to their datasets for other researchers to reuse or to ensure that their results are reproducible. Dataverse, developed by Harvard's Institute for Quantitative Social Science (IQSS), is an open-source web application that allows researchers, journals, publishers and research institutes around the world to house, share, organize and archive their project data.

3ie has shared more than 100 impact evaluation datasets on Dataverse.[8] An advanced search of datasets with the subject: "social science" and keyword: "education" or "school" or "student" or "learning," yield 9 results of impact evaluation datasets available for download. In most cases, do-files also come along with the database for other researchers to replicate and verify the published findings. Recently in late 2019, Innovations for Poverty Action (IPA) and the Abdul Latif Jameel Poverty Action Lab (J-PAL) jointly created Datahub for Field Experiments in Economics and Public Policy (DFEEP) and started housing their project databases on the Dataverse Project.[9] At the time of writing, 149 datasets (at least 14 are education-related) from the two institutes are available for download along with the statistical codes and related documents.

Another good source to obtain the RCT dataset is the World Bank's microdata library.[10] A large proportion of the World Bank's development projects consists of an impact evaluation component; however, it seems not all the impact evaluation databases are shared with the public on the website. Currently, only 158 datasets are accessible.

5. Predicting Student's Success with Educational Data Mining

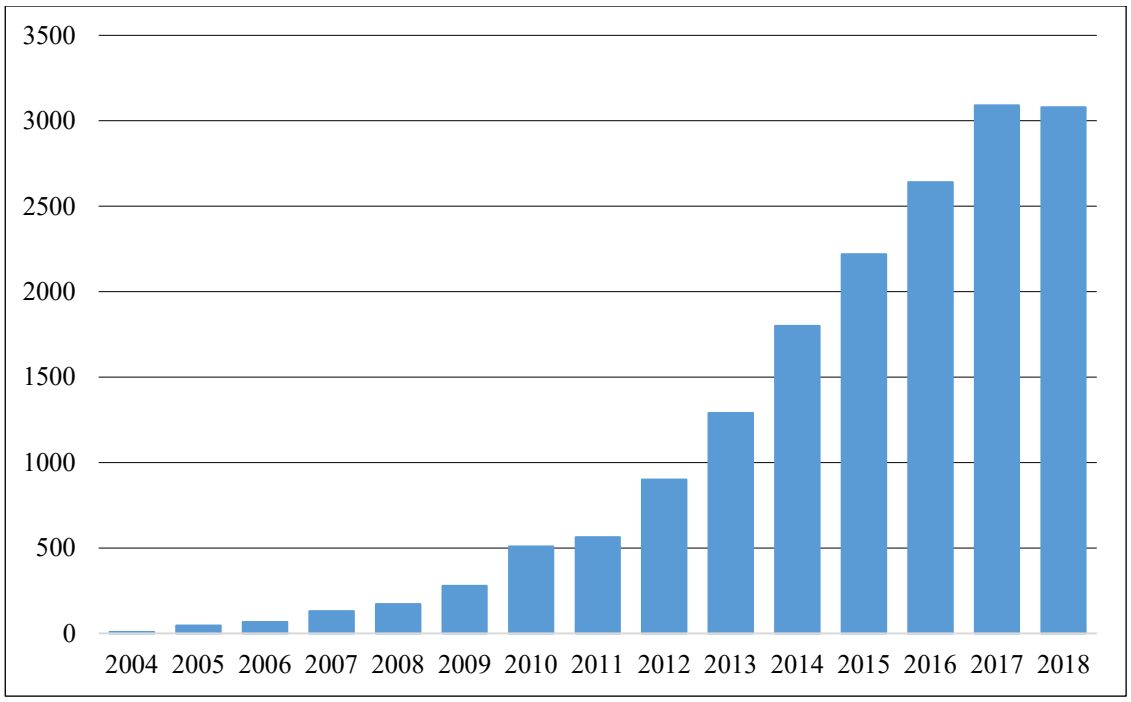
Empirical studies in the field of education, more or less, have tried to identify factors or program interventions that can improve students' learning, expand education access or narrow the gaps between sub-groups of the population. In other words, it attempts to understand what and why it happened. The advancement of technology and computational power have motivated researchers to apply computer science in educational settings. Educational data mining, an emerging interdisciplinary research field of computer science, statistics, and education, is trying to better understand and predict students' success or failure and/or to identify at-risk students in the future

[8] https://dataverse.harvard.edu/dataverse/3ie
[9] https://dataverse.harvard.edu/dataverse/DFEEP
[10] https://microdata.worldbank.org/index.php/catalog/impact_evaluation

using data mining and machine learning techniques (Romero & Ventura, 2013). It is designed to automatically detect meaningful patterns from a large amount of data, for instance, information related to student backgrounds or learning activities. Asif et al. , (2017) used the educational data mining approach to identify poor undergraduate students in Pakistan, so that schools can provide timely warnings and advice. The commonly used methods for prediction include classification, regression, clustering, and feature selection (Bakhshinategh et al., 2018). In Educational Data Mining's educational knowledge discovery process, the decision tree algorithm is one of the most widely used methods due to its accuracy and ease in interpretation (Romero et al., 2013).

A Google Scholar's search using the key term of "educational data mining," suggests that the number of studies on this topic has steadily increased in the last ten years. We found no publication related to educational data mining in 2004, but Google Scholar's search produced more than 3,000 results published in 2017. As a result of the growing interest in educational data mining, the Journal of Educational Data Mining was established in 2009 and two years later in 2011, the International Educational Data Mining Society was founded.

Figure 3. Google Scholar Search Results Using "Educational Data Mining"



Source: Created by the author based on Google Scholar search results

However, a majority of these studies are conducted by researchers in the field of computer science rather than by researchers in the field of education development. This is probably due to the fact that researchers in the field of education development are not trained and not familiar with the machine-learning algorithms.

6. Conclusion

Thanks to global initiatives from international organizations such as the World Bank, UNICEF, OECD and UNESCO, more and more data related to education are publicly available for researchers and academics. Yet, in many countries, particularly in developing countries, openly accessible data are still limited. Data mentioned in the paper are mainly collected by international organizations; however, there are many more education data collected by national statistics, ministries of education, schools, or through specific projects. Unfortunately, most of these data are locally stored and restricted for internal use only. Nevertheless, there have also been efforts from universities and research institutes to make open-source data more accessible to researchers. This data sharing allows other researchers to replicate the analysis and promote transparency in social science research.

Educational Data Mining can be a potential tool in education research in the near future although currently most studies using this approach are done by researchers in the field of computer science without deep knowledge of education development issues. Since the amount of data has increased rapidly in recent years, there should be more collaboration between researchers of both fields using big data in identifying education issues and its solutions.

Acknowledgment

The author is deeply grateful to Prof. Takashi Kamihigashi and Prof. Koji Yamazaki for their ideas and suggestions at the initial stage of the study. The author is also would like to express his thanks to his colleagues at RIEB and GISCS for their comments on the draft.

References

Altinok, N., Angrist, N., & Patrinos, H. A. (2018). *Global Data Set on Education Quality (1965-2015)* (Policy Research Working Papers No. 8314). Washington D.C: World Bank.

Ammermüller, A., Heijke, H., & Wößmann, L. (2005). Schooling Quality in Eastern Europe: Educational Production During Transition. *Economics of Education Review*, *24*(5), 579–599. https://doi.org/10.1016/j.econedurev.2004.08.010

Angrist, J. D., & Pischke, J.-S. (2008). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press.

Asif, R. et al., (2017). Analyzing Undergraduate Students' Performance Using Educational Data Mining. *Computers and Education*, *113*, 177–194. https://doi.org/10.1016/j.compedu.2017.05.007

Bakhshinategh, B., Zaiane, O. R., ElAtia, S., & Ipperciel, D. (2018). Educational Data Mining Applications and Tasks: A Survey of the Last 10 Years. *Education and Information Technologies*, *23*(1), 537–553. https://doi.org/10.1007/s10639-017-9616-z

Canton, E., & Blom, A. (2004). *Can Student Loans Improve Accessibility to Higher Education and Student Performance?: An Impact Study of the Case of SOFES, Mexico* (World Bank Policy Research Working Paper No. 3425). World Bank.

Connelly, R., & Zheng, Z. (2003). Determinants of school enrollment and completion of 10 to 18 year olds in China. *Economics of Education Review*, *22*(4), 379–388.

Connolly, P., Keenan, C., & Urbanska, K. (2018). The Trials of Evidence-based Practice in Education: A Systematic Review of Randomised Controlled Trials in Education Research 1980–2016. *Educational Research*, *60*(3), 276–291. https://doi.org/10.1080/00131881.2018.1493353

Crawfurd, L., Hares, S., Nestour, A., & Minardi, A. (2019). PISA 2018: A Few Reactions to the New Global Education Rankings. Retrieved December 13, 2019, from https://www.cgdev.org/blog/pisa-2018-few-reactions-new-global-education-rankings

Cresswell, J., Schwantner, U., & Waters, C. (2015). *A Review of International Large-Scale Assessments in Education: Assessing Component Skills and Collecting Contextual Data*. *OECD Publishing*. Paris: OECD Publishing. https://doi.org/10.1787/9789264248373-en

Duflo, E., Glennerster, R., & Kremer, M. (2007). Using Randomization in Development Economics Research: A Toolkit. In *Handbook of Development Economics* (Vol. 4, pp. 3895–3962). Elsevier B.V. https://doi.org/10.1016/S1573-4471(07)04061-2

Edmonds, E. (2007). Child Labor. In *Handbook of Development Economics* (Vol. 4, pp. 3607–3709). Elsevier B.V. https://doi.org/10.1016/S1573-4471(07)04057-0

Education Policy and Data Center. (2009). *How (Well) is Education Measured in Household Surveys ? A Comparative Analysis of the Education Modules in 30 Household Surveys from 1996–2005* (IHSN Working Paper No. 002).

Fausset, R. (2014, September 29). Trial Opens in Atlanta School Cheating Scandal. *New York Times*. Retrieved from https://www.nytimes.com/2014/09/30/us/racketeering-trial-opens-in-altanta-schools-cheating-scandal.html

Fuchs, T., & Wößmann, L. (2008). What Accounts for International Differences in Student Performance? A Re-examination Using PISA Data. In *The Economics of Education and Training* (pp. 209–240). Physica-Verlag HD.

Gaddah, M., Munro, A., & Quartey, P. (2015). The Rich or the Poor: Who Gains from Public Education Spending in Ghana? *International Journal of Social Economics*, *42*(2), 112–131. https://doi.org/10.1108/IJSE-11-2013-0269

Glewwe, P., Hanushek, E., Humpage, S., & Ravina, R. (2011). *School Resources and Educational Outcomes in Developing Countries: A Review of the Literature from 1990 to 2010* (NBER Working Paper Series No. 17554). Cambridge, MA: National Bureau of Economic Research.

Guarcello, L., Lyon, S., & Valdivia, C. (2015). *Evolution of the Relationship Between Child Labour and Education Since 2000: Evidence from 19 Developing Countries*. Understanding Children's Work (UCW).

Hanushek, E., & Woessmann, L. (2011). The Economics of International Differences in Educational Achievement. In *Handbook of the Economics of Education* (Vol. 3, pp. 89–200). Elsevier B.V. https://doi.org/10.1016/B978-0-444-53429-3.00002-8

Iwasawa, M., Inada, M., & Fukui, S. (2014). *How Migrant Heterogeneity Influences the Effect of Remittances on Educational Expenditure: Empirical Evidence from the Cambodian Socio-Economic Survey* (No. August 2014). *Kyoto University*. Kyoto.

Jacob, B. A. (2005). Accountability, Incentives and Behavior: The impact of High-Stakes Testing in the Chicago Public Schools. *Journal of Public Economics*, 761–796. https://doi.org/10.1016/j.jpubeco.2004.08.004

Lassibille, G., & Tan, J.-P. (2007). Benefit Incidence Analysis in Education. *Journal of Education Finance*, *33*(2), 170–182.

Leste, A. (2005). Streaming in Seychelles: From SACMEQ Research to Policy Reform. In *SACMEQ Research Conference*. Paris: UNESCO.

Mani, S., Hoddinott, J., & Strauss, J. (2013). Determinants of Schooling: Empirical Evidence from Rural Ethiopia. *Journal of African Economies*, *22*(5), 693–731. https://doi.org/10.1093/jae/ejt007

Mansuri, G. (2006). *Migration, Sex Bias, and Child Growth in Rural Pakistan* (Policy Research Working Papers No. 3946). Washington D.C: World Bank.

Nakata, S. (2013). *Equitable Access to Higher Education in Vietnam: Effects of Government's Financial Support for Low Income Students*. Kobe University.

Pritchett, L. (2013). *The Rebirth of Education: Schooling Ain't Learning*. Washington D.C: Center for Global Development.

Psacharopoulos, G. (1994). Returns to Investments in Education: A Global Update. *World Development*, *22*(9), 1325–1343. https://doi.org/10.1016/0305-750X(94)90007-8

Psacharopoulos, G., & Patrinos, H. A. (2018). Returns to Investment in Education: A Decennial Review of the Global Literature. *Education Economics*, *26*(5), 445–458. https://doi.org/10.1080/09645292.2018.1484426

Rapoport, H., & Docquier, F. (2006). The Economics of Migrants' Remittances. In *Handbook of the Economics of Giving, Altruism and Reciprocity* (pp. 1135–1198).

Raudonyte, I. (2019). *Use of Learning Assessment Data in Education Policy-Making*. Paris: IIEP/UNESCO.

Romero et al., (2013). Data Mining in Education. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *3*(1), 12–27. https://doi.org/10.1002/widm.1075

Sánchez, A., & Singh, A. (2018). Accessing Higher Education in Developing Countries: Panel data Analysis from India, Peru, and Vietnam. *World Development*, *109*, 261–278.

Snilstveit, B., Stevenson, J., Phillips, D., & Vojtkova, M. (2015). *Interventions for Improving Learning Outcomes and Access to Education in Low- and Middle-Income Countries: A Systematic Review* (3ie Systematic Review No. 24). London: International Initiative for Impact Evaluation (3ie).

Tansel, A. (2002). Determinants of School Attainment of Boys and Girls in Turkey: Individual, Household and Community Factors. *Economics of Education Review*, *21*(5), 455–470. https://doi.org/10.1016/S0272-7757(01)00028-0

Thaler, R., & Sunstein, C. (2008). *Nudge: Improving Decisions About Health, Wealth, and Happiness*. New Haven: Yale University Press.

Tobin, M., Lietz, P., Nugroho, D., Vivekanandan, R., & Nyamkhuu, T. (2015). *Using Large-scale Assessments of Students ' Learning to Inform Education Policy: Insights from the Asia-Pacific Region*. *Australian Council for Educational Research*. Melbourne: Australian Council for Educational Research.

UIS. (2004). *Guide to the Analysis and Use of Household Survey and Census Education Data*. Montreal, Quebec: UNESCO Institute for Statistics.

UIS. (2017). *The Data Revolution in Education* (UIS Information Paper No. 39). Montreal, Quebec: UNESCO Institute for Statistics. https://doi.org/10.15220/978-92-9189-213-6-en

UIS. (2019). *New Methodology Shows that 258 Million Children, Adolescents and Youth Are Out of School*. Montreal, Quebec: UNESCO Institute for Statistics. Retrieved from http://uis.unesco.org/sites/default/files/documents/new-methodology-shows-258-million-children-adolescents-and-youth-are-out-school.pdf

UNESCO. (2013). *The Global Learning Crisis: Why Every Child Deserves a Quality Education*. *EFA Global Monitoring Report*. Paris: UNESCO.

UNESCO. (2015). *Education for All 2000–2015: Achievements and Challenges*. *EFA Global Monitoring Report*. Paris: UNESCO.

World Bank. (2018). *Learning to Realize Education's Promise*. *World Development Report*. Washington, DC: World Bank.

Wößmann, L., & West, M. (2006). Class-size effects in school systems around the world: Evidence from between-grade variation in TIMSS. *European Economic Review*, *50*(3), 695–736. https://doi.org/10.1016/j.euroecorev.2004.11.005

Yadav, S. K., Bharadwaj, B., & Pal, S. (2012). Data Mining Applications: A Comparative Study for Predicting Student's Performance. *International Journal of Innovative Technology & Creative Engineering*, *1*(12), 13–19.

Annex

Table A1. List of Countries Conducting DHS, MISC or LSMS

Region	Country	DHS	MISC	LSMS
Sub-Saharan Africa	Angola		✓	✓
	Benin	✓	✓	
	Botswana	✓	✓	
	Burkina Faso	✓	✓	✓
	Burundi	✓	✓	
	Cameroon	✓	✓	
	Cape Verde	✓		
	Central African Republic	✓	✓	
	Chad	✓	✓	
	Comoros	✓	✓	
	Democratic Republic of the Congo	✓	✓	
	Cote d'Ivoire	✓	✓	✓
	Djibouti		✓	
	Equatorial Guinea	✓	✓	
	Eritrea	✓		
	Ethiopia	✓	✓	✓
	Gabon	✓	✓	
	Gambia	✓	✓	
	Ghana	✓	✓	✓
	Guinea	✓	✓	
	Guinea-Bissau		✓	
	Kenya	✓	✓	
	Lesotho	✓	✓	
	Liberia	✓	✓	
	Madagascar	✓	✓	
	Malawi	✓	✓	✓
	Mali	✓	✓	✓
	Mauritania	✓	✓	
	Mozambique	✓	✓	
	Namibia	✓		
	Niger	✓	✓	✓
	Nigeria	✓	✓	✓
	Rwanda	✓	✓	
	Sao Tome and Principe	✓	✓	
	Senegal	✓	✓	
	Sierra Leone	✓	✓	
	Somalia		✓	
	South Africa	✓		✓
	South Sudan		✓	
	Sudan	✓	✓	
	Swaziland	✓	✓	
	Tanzania	✓	✓	✓

Region	Country	DHS	MISC	LSMS
	Togo	✓	✓	
	Uganda	✓		✓
	Zambia	✓	✓	
	Zimbabwe	✓	✓	
Northern Africa	Algeria		✓	
	Egypt	✓	✓	
	Libya		✓	
	Morocco	✓		
	Tunisia	✓	✓	
Central Asia	Kazakhstan	✓	✓	✓
	Kyrgyzstan	✓	✓	✓
	Tajikistan	✓	✓	✓
	Turkmenistan	✓	✓	
	Uzbekistan	✓	✓	
Eastern Asia	China		✓	✓
	Korea, Democratic People's Republic of		✓	
	Mongolia		✓	
South-East Asia				
	Cambodia	✓		
	Indonesia	✓	✓	
	Lao People's Democratic Republic	✓	✓	
	Myanmar	✓	✓	
	Philippines	✓	✓	
	Thailand	✓	✓	
	Timor-Leste	✓		✓
	Vietnam	✓	✓	✓
Southern Asia	Afghanistan	✓	✓	
	Bangladesh	✓	✓	
	Bhutan		✓	
	India	✓	✓	✓
	Iran		✓	
	Maldives	✓	✓	
	Nepal	✓	✓	✓
	Pakistan	✓	✓	✓
	Sri Lanka	✓		
Western Asia	Armenia	✓		✓
	Azerbaijan	✓	✓	✓
	Bahrain		✓	
	Georgia		✓	
	Iraq		✓	✓

Region	Country	DHS	MISC	LSMS
	Jordan	✓		
	Lebanon		✓	
	Oman		✓	
	Qatar		✓	
	State of Palestine		✓	
	Syrian Arab Republic		✓	
	Turkey	✓	✓	
	West Bank/Gaza	✓		
	Yemen	✓	✓	
Latin America and the Caribbean	Argentina		✓	
	Barbados		✓	
	Belize		✓	
	Bolivia	✓	✓	
	Brazil	✓		✓
	Colombia	✓		
	Costa Rica		✓	
	Cuba		✓	
	Dominican Republic	✓	✓	
	Ecuador	✓		✓
	El Salvador	✓	✓	
	Guatemala	✓		✓
	Guyana	✓	✓	✓
	Haiti	✓		
	Honduras	✓	✓	
	Jamaica		✓	✓
	Mexico	✓	✓	
	Nicaragua	✓		✓
	Panama		✓	✓
	Paraguay	✓	✓	
	Peru	✓		✓
	Saint Lucia		✓	
	Suriname		✓	
	Trinidad and Tobago	✓	✓	
	Turks and Caicos Islands		✓	
	Uruguay		✓	
	Venezuela		✓	
Oceania	Fiji		✓	
	Kiribati		✓	
	Nauru		✓	
	Papua New Guinea	✓		
	Samoa	✓	✓	
	Tonga		✓	
	Tuvalu		✓	
	Vanuatu		✓	

Region	Country	DHS	MISC	LSMS
Eastern Europe	Albania	✓	✓	✓
	Belarus		✓	
	Bosnia and Herzegovina		✓	✓
	Bulgaria			✓
	Croatia		✓	
	Kosovo		✓	✓
	Moldova	✓	✓	
	Montenegro		✓	✓
	North Macedonia		✓	
	Serbia		✓	✓
	Ukraine	✓	✓	
	Yugoslavia		✓	

Sources: Created by the author based on databases from DHS, MICS, and LSMS.

Table A2. List of Selected Learning Assessments

Assessment	Year	Organization	Locations	Subjects
International (School-based)				
PISA	2000, 2003, 2006, 2009, 2012, 2015, 2018	OECD	International	R, M, S
PISA-D	2018	OECD	Low-and Middle Income Countries	R, M, S
PIRLS	2001, 2006, 2011, 2016	IEA	International	R
TIMSS	1995, 1999, 2003, 2007, 2011, 2015	IEA	International	M, S
EGRA	Since 2007	RTI and USAID	International	R
EGMA	Since 2009	RTI and USAID	International	M
Regional (School-based)				
SACMEQ	1995-1999, 2000-2004, 2006-2011, 2012-204	SEACMEQ Consortium	Southern and Eastern Africa	R, M
PASEC	1993-2011, 2002-2012, 2014	CONFEMEN	Francophone Sub-Saharan Africa	R, M
LLECE	1997, 2006, 2013	UNESCO	Latin America	R, M, S, W
PILNA	2012	EQAP and UNESCO	Pacific Island	L, N
SEA-PLM	2019	SEAMEO and UNICEF	Southeast Asia	R, M, W, GC
Household-based				
ASER	Annually since 2005	Pratham	India, Pakistan	R, N
UWEZO	Annually since 2009	Twaweza	Kenya, Tanzania, Uganda	L, N
PIAAC	2011-2012, 2014-2015, 2017	OECD	International	L, N, PS
STEP	2011, 2012, 2014	World Bank	Low-and Middle Income Countries	R
LAMP	Since 2003	UNESCO, UIS	Low-and Middle Income Countries	L, N

- Note: R=Reading, M=Mathematics, S=Science, W=Writing, L=Literacy, N=Numeracy, GC=Global Citizenship, PS=Problem Solving
- Source: Created by the author based on Altinok et al., (2018), Cresswell, Schwantner, & Waters (2015), Raudonyte (2019), and Organizers' websites.