UNESCO Conference on Education Data and Statistics
Session 1
7–9 February 2024

1 UIS/EDS/10

INTEGRATION OF STATISTICS: CHALLENGES AND SOLUTIONS FORWARD

FEBRUARY 2024

2024 CONFERENCE ON
EDUCATION DATA
AND **STATISTICS**

1. INTRODUCTION

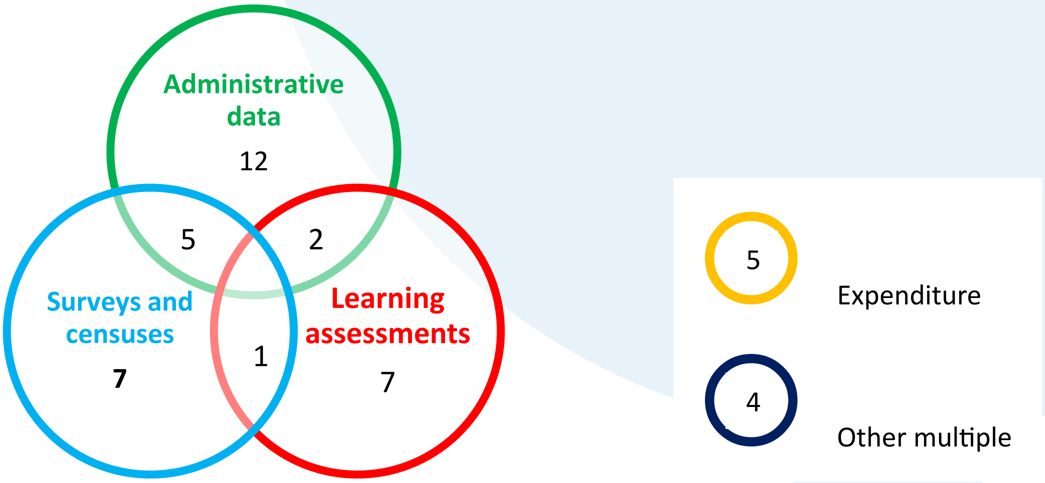
The SDG monitoring framework entails a substantial level of ambition and requires a corresponding amount of innovation. Calls for a post-2015 data revolution reflected this sense of urgency for new ideas. One key suggestion from 2014 was that ‘the more data can be combined, the more useful they are’ (Independent Expert Advisory Group on the Data Revolution for Sustainable Development, 2014).

There are several education indicators that potentially draw on multiple data sources or types of data sources:

- Some indicators can be informed by multiple sources of the same type. These could – due to differences in methodologies, objectives or circumstances at the point of data collection – produce results that are not fully comparable without further analysis. Examples are learning outcome indicators based on different assessments and wealth parity indices indicators based on different surveys (which may measure wealth in different ways).
- Some indicators can be informed by different types of sources (Figure 1). One example is the out-of-school rate, which can rely on both administrative and survey data. Another example is teacher continuous professional development, which can draw on both administrative and learning assessment data. Some countries may opt for only one source of data instead of the other.

The problem that education statisticians are therefore increasingly called to solve is how to incorporate multiple data sources or types of data sources in the estimation of indicators.

Figure 1. Distribution of SDG 4 global and thematic indicators, by potential data source



Even before the SDGs, multiple data sources and types of data sources had similarly challenged other sectors in producing estimations of selected indicators. For example:

- The need to use multiple surveys with different methodologies (as well as to address data gaps) to estimate malnutrition indicators, such as wasting and stunting, led to the establishment of the Joint Child Malnutrition Estimates inter-agency group in 2011 (UNICEF et al., 2023).
- The need to use multiple administrative and survey data sources to estimate health indicators led to the establishment of the UN Inter-agency Group for Child Mortality Estimation and the adoption of a model to generate annual estimates for under-5 mortality (Alkema and New, 2014) and the establishment of the UN Maternal Mortality Estimation Interagency Group (Alkema et al., 2016).

Similar steps have recently been taken in education. But opportunities come with challenges. This paper outlines these steps, emerging challenges and a potential forward agenda for more effective data integration.

2. PROGRESS ACHIEVED

Multiple data sources and types of data sources have been recently used to estimate two education indicators: the completion rate (indicator 4.1.2) and the out-of-school rate (indicator 4.1.4).

2.1. A completion rate model

The Inter-agency and Expert Group on SDG Indicators adopted the completion rate at three levels of education (primary, lower secondary and upper secondary) as SDG global indicator 4.1.2, one of only six successful proposals out of more than 200 made during the 2020 Comprehensive Review of the SDG Monitoring Framework. Indicator 4.1.2 is defined as the 'percentage of a cohort of children or young people aged 3–5 years above the intended age for the last grade of each level of education who have completed that grade'. The completion rate is a 'flow' measure of attainment, which recognizes late enrolment and high repetition in many poorer countries that lead many children to reach the end of each education cycle several years after the official graduation age.

Combining multiple survey data sources can tackle some problems, such as infrequent survey cycles (every three to five years) and a variety of sampling and non-sampling errors that generate conflicting information between different surveys in the same country. However, surveys also have advantages over administrative data, such as the better recording of age information and the universal coverage of education programmes. It is also possible to use retrospective information to reconstruct the historical completion rates of older cohorts and not be limited just to information about the most recent cohort.

The developed model is a Bayesian hierarchical model inspired by the approach used to estimate health indicators but adapted to the education context (Dharamshi et al., 2022). It estimates an underlying trend in target values and shares information on parameter scaling across countries. Late completion is explicitly modelled by specifying the magnitude of the delay as a function of age. Age misreporting concerns are also addressed.

Such adjustments permit the model to consolidate survey data into a smooth underlying completion rate trend from which the estimated true annual completion rates for each country can be extracted. By addressing the various data quality concerns associated with survey data, these estimates are also less sensitive to individual surveys, the year in which they were conducted, and the type of survey that happens to be the latest available in a given country.

Point estimates continue to be reported for combinations of individual countries and survey years in the UIS database. But the UIS also provides the model estimates alongside these point estimates, while the Technical Cooperation Group (TCG) on SDG 4 Indicators has approved the use of the model estimates for regional and global aggregates for the SDG database.

The completion rate is the survey-based counterpart of an administrative data–based indicator, the gross intake rate to the last grade of school. A potential future extension of the model could be to combine survey and administrative data (as with the out-of-school rate model, which is presented next).

2.2. An out-of-school rate model

The out-of-school rate is the 'proportion of children and young people in the official age range for the given level of education who are not enrolled in pre-primary, primary, secondary or higher levels of education' (UIS, 2021). It was the flagship indicator in 2000–15 under the Education for All agenda and the Millennium Development Goals. The need for a methodology that combines data sources to estimate out-of-school rates was recognized 20 years ago, when it was acknowledged that 'some sort of composite approach may be needed for estimating time series and producing estimates for the most recent year' (UIS and UNICEF, 2005).

In the absence of such a 'composite approach', the calculation of out-of-school rates has been based on enrolment records from school censuses. However, using administrative data has three challenges in poorer countries with high out-of-school rates. First, enrolment records are often incomplete, inaccurate or missing altogether. Second, estimates need to combine enrolment counts with a measure of the population, which comes from a different and often inconsistent source. The quality of single-age population estimates, required to calculate out-of-school rates accurately, is often not high, leading to jumps in the indicator time series – and sometimes to more children being recorded as enrolled than the number of children of that age group. Third, with low birth registration rates, the capacity of schools to record student age accurately is limited.

In recent years, many of these countries have carried out household surveys which, despite their own weaknesses, can help fill some gaps and address challenges related to age and population. A Bayesian hierarchical cohort-based model was accordingly developed to estimate out-of-school rates using multiple data sources (UIS and GEM

Report, 2022). The model mirrors the natural progression of students through the school cycle. Data from both administrative and survey sources are reconciled, recognizing the differences in the nature and generation of data of these two types, while sharing information about bias and variance across countries.

The model introduces some new key ideas. First, it uses a cohort approach to link out-of-school rates, similar to demographic modelling of population processes. However, it is due to the lack of reliable data that the risks of late entry, dropout, repetition and other relevant education transitions are not estimated; instead, net grade-to-grade changes are estimated flexibly to smoothen the underlying out-of-school rate cohort curves. Second, the model accepts cases where there are more children in school than children of a particular age, but at the same time constrains out-of-school rates to be between 0 and 100% in order to allow such information to be used. Third, the model shifts the focus from out-of-school rates by education level to out-of-school rates by age, as students enter and exit school at every age.

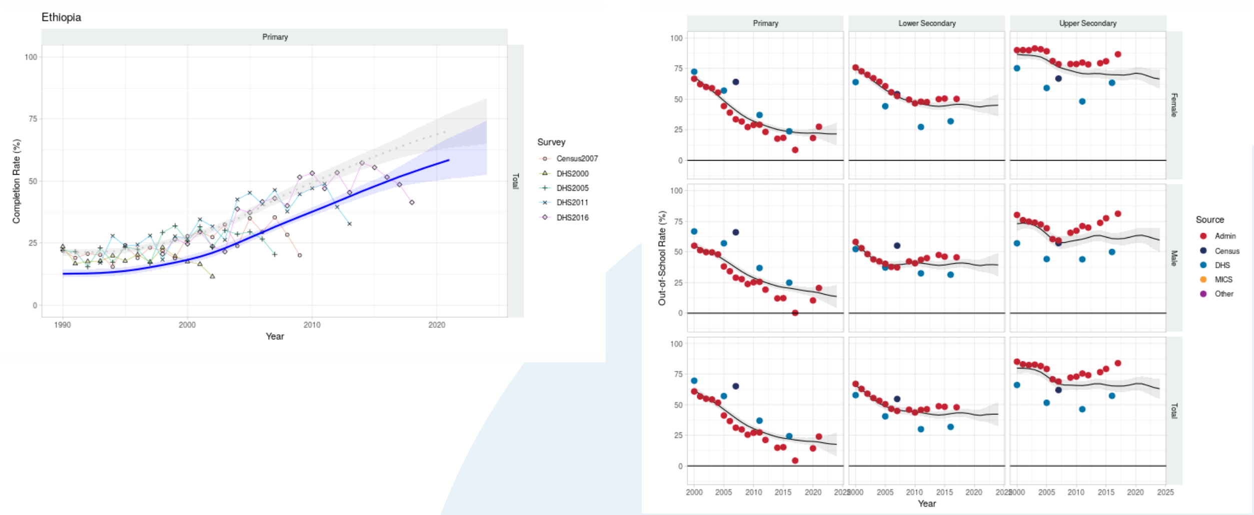
The results of this model were reported for the first time in September 2022, including for many countries that have not had administrative data on out-of-school rates for many years. While administrative data estimates remain the officially reported national data in the UIS database, as with the completion rate, the UIS also provides model estimates for individual countries, while model estimates are the preferred source for regional and global aggregates. An update in September 2023 estimated there were 250 million children, adolescents and youth in 2022.

Despite the efficient use of all available information, the model is also subject to its own challenges. The main one is that, as the model is driven by a cohort approach and is characterized by a degree of inertia, it cannot easily incorporate a major impact on school attendance from one year to the next caused by emergencies. Other challenges relate to consistency and are discussed in Section 4.

2.3. Visualizing the completion and out-of-school rate model estimates

Results of both models are available on the VIEW website, maintained by the Global Education Monitoring Report and the UNESCO Institute for Statistics and updated twice a year. The site visualizes the input data, inviting a closer inspection of the statistical problems that are being addressed. In the case of the completion rate, the visualization showcases the challenge of late enrolment. A large number of children complete each level of education even after 3 to 5 years above the intended age for the last grade. The model, in fact, estimates not only the official completion rate but also an 'ultimate' completion rate, which includes those who finish school up to 8 years late (**Figure 2a**). In the case of the out-of-school rate, the visualization highlights each source of data with a separate colour, which helps identify the source of discrepancies (**Figure 2b**).

Figure 2. Visualization of input data and model estimates for two education indicators



a. Primary completion rate, Ethiopia

b. Out-of-school rate, by age group and sex, Ethiopia

Source: https://education-estimates.org/.

3. CHALLENGES ASSOCIATED WITH ESTIMATING INDICATORS USING MULTIPLE DATA SOURCES

Recent efforts to combine multiple information sources to estimate and report SDG 4 indicators are important steps taken by UNESCO to fulfil its responsibility for informing the international community through efficient use of multiple sources data. Nevertheless, continuing efforts to integrate data also need to ensure that best practice is followed, that countries' needs are served, and that consistent estimates are calculated. This section briefly outlines these challenges.

3.1. Ensure best practice in reporting estimates based on multiple data sources

Education data are rarely available for every population and year, while data from multiple sources and differences in measurement methodologies may result in inconsistent and non-comparable estimates. Despite a clear commitment that countries must lead SDG reporting, the use of models may be unavoidable to get a sense of how key indicators evolve. Yet the data and methodologies used to produce estimates often have features or assumptions that affect their interpretation. Accurate interpretation and responsible use of estimates requires understanding on what the data used to base estimates were based (how they were identified, accessed and included; what is their quality) and of the complex methods used to derive the estimates.

3.2. Ensure country participation and ownership in the generation of estimates

Even if best practice is followed, further protocols are needed to ensure that countries, which as mentioned above are expected to lead SDG reporting, can participate in and engage with estimation processes. Such protocols are necessary to facilitate country ownership of generated estimates and to help develop the capacity of national statisticians and experts to contribute to the process improvements. Currently, there is no systematic mechanism for countries to seek clarifications, understand the methods underpinning the estimates and contest results that contradict their own understanding of the actual situation, as well as to proactively contribute data sources and ideas for the development of the models.

3.3. Ensure consistency between out-of-school and completion rate estimates

The completion and out-of-school models aim to solve inconsistencies between and within multiple data sources and types of sources. Each of these estimates is internally consistent: completion rate estimates for a given year are consistent with past estimates and out-of-school rate estimates at a given age are consistent with a cohort's out-of-school rates at previous ages. But there are other consistency issues to be resolved. First, rates for females and males are currently estimated independently of each other and can arbitrarily diverge. Second, completion and out-of-school rates are estimated independently, even though they are not independent of each other: every child completing school must have spent a certain number of years enrolled, putting a constraint on how many children can have been out of school.

4. AGENDA FORWARD

The challenges described in this paper point to potential solutions that can build on the progress made so far. These solutions also aim to strengthen the community of practice of education statisticians.

4.1. Formalize good practice for reporting estimates

Faced with similar challenges related to the availability of multiple sources, the international health statistics community, under the leadership of the World Health Organization, issued the Guidelines for Accurate and Transparent Health Estimates Reporting (GATHER) to define best reporting practices for studies that estimate indicators using multiple data sources. The process involved a working group that reviewed existing reporting guidelines, generated a list of potential reporting items, received feedback from a broader community of researchers and users of estimates, and drafted an essential checklist of 18 items that should be reported whenever estimates are published, in order to serve the needs of decision makers and researchers (Stevens et al., 2016).

The GATHER checklist of information that should be included in reports of global estimates is organized into four sections: (1) objectives and funding, (2) data sources, (3) data analysis, and (4) results and discussion. The intention is that information about data sources and analysis methods, including key assumptions and limitations, are presented in a way that is accessible without advanced training in statistics. Explanations of how new estimates compare to previously published estimates, and why they differ, should be provided.

The guidelines also aim to promote open access to data inputs and source code, as full documentation increases the value of research. Data underlying estimates should be accessible online, which might require additional resources for documentation and archiving. Exceptions include cases when this is not possible, such as third-party ownership. Sharing source code also involves an investment of resources, especially if the code is fully documented and available online for off-the-shelf use. As a minimum, key segments of code should be shared but researchers should not be held responsible for providing user support.

Authors should also report a measure of the uncertainty associated with estimates, such as uncertainty intervals. Global estimates are affected by multiple sources of error, such as measurement error during data collection, inability to obtain a truly random sample, errors in adjusting input data for sources of bias, and the use of a model to calculate estimates. Users of these estimates should be informed about their overall uncertainty. The reporting guidelines are designed to be flexible enough to guide reporting of estimates regardless of the underlying data availability and the complexity of the statistical methods.

It is proposed that:

- The TCG initiates a process to discuss emerging issues from using models to estimate SDG 4 indicators.
- GATHER best reporting practices should be adapted from the health to the education sector.

4.2. Support country participation and ownership in estimates

The development of models to estimate education indicators has been initiated by UNESCO. However, now that these examples available, it is time for countries to review the results in a systematic way, familiarize themselves with the rationale and implications, identify errors and seek clarifications, contribute ideas to potential areas of model development, and provide additional and up-to-date data sources. Familiarizing ministries of education and the expert community with estimate-based SDG 4 indicators as a new way of monitoring progress requires extensive communication.

In international health statistics, such as in mortality indicator estimates, there are processes that aim to strengthen national capacity in collecting data, understanding estimation techniques and interpreting results. Regional workshops have been used to train participants from different countries in techniques and modelling methods underlying the estimates. Experts have been sent to countries to conduct training on child mortality estimation. As part of a data review process, WHO and UNICEF through their field offices consult governments, which provide feedback on the plausibility of estimates and the validity of underlying data. These efforts are built on the principle that indicator estimation is not simply an academic exercise but a fundamental part of effective policy and programming.

It is proposed that similar steps are also taken in education statistics, whereby:

- The UNESCO Institute for Statistics includes model estimates in the agenda of its existing programme of regional capacity development workshops to familiarize countries with the methods.
- The development of an inventory of surveys (one of the solutions proposed in the conference paper on household surveys) also serves data integration, ensuring countries are involved in the data inputs used.

4.3. Develop a joint model of out-of-school and completion rates

The Global Education Monitoring Report and the UNESCO Institute for Statistics are currently working to develop a computationally feasible model that integrates and ensures the consistency of completion and out-of-school rate estimates with each other (and potentially estimates of other related indicators), disaggregated by sex.

A joint model can assess discrepancies in a systematic way and help adjust completion and out-of-school rates in accordance with each other: joint estimates are expected to be more accurate than independent estimates. To date, completion rate estimates (which only rely on survey sources) have been more precise than out-of-school rate estimates (which rely on survey and administrative sources). This implies that out-of-school rates are more likely to change substantially in some countries in the context of a joint model.

Completion and out-of-school rate estimates are related to each other via age-grade progression, i.e. age- and grade-specific entry, repetition and dropout behaviour. In theory, joint estimates could be improved by making use of data on these flows; completion and out-of-school rate estimates could likewise serve to inform progression estimates. In practice, neither ambition is likely to be feasible. The coverage and quality of age-grade progression data at the level of single years of age or individual grades is poor. A joint model should therefore rely only on weak assumptions regarding progression, occasionally using relevant pieces of information, for example, on automatic promotion and zero repetition policies. Neither does it identify nor require all the elements of a comprehensive progression model.

It is proposed that a joint completion and out-of-school model be developed, taking into account that:

- All schooling levels should be modelled jointly to properly account for the effects of late entry to school.
- The joint model needs to be simple enough to keep the computational burden manageable.
- The model should be structured in terms of the enrolment implications of conditioning on the shares of different groups of completers (or non-completers) and those who never entered school.

- These enrolment implications of completion patterns can be modelled with various degrees of flexibility (e.g. a single late-entry effect vs different late entry rates of different groups of completers; constant levels of late entry vs changing late entry trends over time).
- To the extent possible, the possibility that this model also helps estimate other related indicators (e.g. enrolment ratios) should be explored.

4.4. Develop models to estimate other indicators that rely on multiple data sources

While the completion and out-of-school rates have been prioritized for model development, these are only two of a larger set of indicators that could benefit from the systematic use of multiple data sources and types of data sources. It is proposed that the TCG explores the possibility of developing models to estimate other SDG 4 indicators, which can draw on:

- Multiple data sources, including:
  - Percentage of children over-age for grade (indicator 4.1.5)
  - Participation rate of youth and adults in formal and non-formal education and training in the previous 12 months, by sex (indicator 4.3.1)
- Multiple types of data sources, including:
  - Participation rate in organized learning (one year before the official primary entry age), by sex (indicator 4.2.2)
  - Gross early childhood education enrolment ratio in (a) pre-primary education and (b) and early childhood educational development (indicator 4.2.4)
  - Gross enrolment ratio for tertiary education by sex (indicator 4.3.2).

The youth and adult literacy rates (indicator 4.6.2) also fall under the latter category. While a model already exists for this indicator, a review and potential extension may be warranted.

REFERENCES

Alkema L. et al. (2016). Global, regional, and national levels and trends in maternal mortality between 1990 and 2015, with scenario-based projections to 2030: a systematic analysis by the UN Maternal Mortality Estimation Inter-Agency Group. *Lancet*. 387 (10017), pp.462-474. https://www.thelancet.com/journals/lancet/article/PIIS01406736(15)00838-7/fulltext

Alkema L. and New, J. R. (2014). Global estimation of child mortality using a Bayesian B-spline bias-reduction model. *Annals of Applied Statistics* 8 (4), pp. 2122-2149. https://www.jstor.org/stable/24522377

Dharamshi, A., Barakat, B., Alkema L., and Antoninis M. (2022) A Bayesian model for estimating Sustainable Development Goal indicator 4.1.2: School completion rates. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 71 (5), pp. 1822–1864. https://academic.oup.com/jrsssc/article/71/5/1822/7073267

Stevens G. et al. (2016) Guidelines for Accurate and Transparent Health Estimates Reporting: the GATHER statement. *Lancet* 388 (10062) E19-E23. https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(16)30388-9/fulltext

UIS and GEM Report (2022). *A Bayesian cohort model for estimating SDG indicator 4.1.4: Out-of-school rates.* Paris: UNESCO. https://www.unesco.org/gem-report/sites/default/files/medias/fichiers/2022/08/OOS_Proposal.pdf

UNICEF, WHO and World Bank (2023). Levels and trends in child malnutrition: UNICEF / WHO / World Bank Group Joint Child Malnutrition Estimates: Key findings of the 2023 edition. New York: United Nations Children's Fund and World Health Organization. https://iris.who.int/bitstream/handle/10665/368038/9789240073791-eng.pdf?sequence=1

UIS (2021). Metadata 4.1.4 Out-of-school rate (1 year before primary, primary education, lower secondary education, upper secondary education). Montreal: UNESCO Institute for Statistics. https://tcg.uis.unesco.org/wp-content/uploads/sites/4/2021/09/Metadata-4.1.4.pdf

UIS and UNICEF (2005). Children out of school: Measuring exclusion from primary education. Montreal: UNESCO Institute for Statistics. https://uis.unesco.org/sites/default/files/documents/children-out-of-school-measuring-exclusion-from-primary-education-en_0.pdf

WEBSITES

VIEW. Visualizing Indicators of Education for the World. Global Education Monitoring Report and UNESCO Institute for Statistics. https://education-estimates.org/

WHO. Guidelines for Accurate and Transparent Health Estimates Reporting (GATHER). World Health Organization. https://www.who.int/data/gather/statement/