International Comparison Program

[07.01]

Towards an Output Approach to Estimate the Value of Education Services in Developing and Transitional Countries

3rd Technical Advisory Group Meeting
June 10-11, 2010

Paris, France

Table of Contents

Towards an Output Approach to Estimate the Value of Education Services in Developing and Transitional Countries

Education Policy and Data Center[1]

1. Introduction – Towards an Output Approach

At the request of the International Comparison Program (ICP) of the World Bank and based on an earlier concept paper provided to the ICP (Pigozzi et.al., 2009), the Education Policy and Data Center has explored refinements for calculating education purchasing power parities for the developing and transitional countries covered by ICP. Our report is divided into two parts. Part I provided a review of the literature on purchasing power parities, and explained why education output was seen as too complex and intangible to measure, and hence was considered resistant to PPP conversions. For these reasons the ICP has used an input-based approach to PPP conversions for education, while recognizing that an output-oriented approach would, if feasible, be a closer, more direct measure of the value of education services. Meanwhile, Eurostat, which calculates the PPPs for OECD countries, is moving towards an output-oriented approach based largely on enrollments and learning scores. This paper, Part II, investigates what would be the necessary components of an output approach to measure the value of education services for the countries covered by the ICP, and proposes a conceptual model for the quantity and quality adjustments to the education PPPs. Further, this paper suggests a roadmap to test the approach and its feasibility in the field, with real data.

The paper is organized in the following manner:

A. Major Components of an Output-Oriented Approach

    2. Measuring the Quantity of Education Services: the Number of

    3. Measuring the Quality of Education Services – how much pupils have learned

B. Critical Adjustments to Quantity and Quality

    4. Time Spent Learning

    5. Non-School Factors

    6. Repetition, Dropout and Length of Schooling

    7. Stratification -- Accounting for Different Types and Delivery Modes of Education Services

[1] The EPDC is a project of the Academy for Educational Development, Washington DC, and is funded by USAID. For more information, visit www.epdc.org.

C. Putting the Components and Adjustments Together

8. Overview of the Proposed Conceptual Model for Output Approach

9. Data Needs, Data Available, Local and External Expertise Required, Anticipated Coverage

10. Roadmap for implementing the pilot phase

2. Measuring the Quantity of Education Services: the Number of Pupils

The primary determinant of the quantity of education services is the number of pupils enrolled in an education system – the recipients of these services. Many ICP countries have an administrative system in place to take an annual (or semi-annual) census of all pupils in primary, secondary and tertiary schools, as well as in preschools, called the Education Management Information System (EMIS).

EMIS pupil counts and data verification

The EMIS information is collected typically by school headmasters and/or teachers who fill in standard questionnaires on the number of pupils and teachers at their school (disaggregating by relevant categories, such as sex, grade level, or training and experience), availability of instructional materials, the number of classrooms and facilities, and periodically, school finances including fees. This information is channeled up the organizational hierarchy, and pooled at the national level.

The EMIS pupil counts, like most data, are not perfect, and need to be verified for accuracy and reliability. In some cases the EMIS information may conflict with other sources of data: for example, attendance rates measured by household surveys can be more than 10% different from enrollment rates measured by the EMIS systems. In other cases, the available counts are out of date, and in some countries, do not exist at all. Further, schools outside mainstream schools, such as community schools, part-time and specialized education (vocational and professional training, arts, sports, etc.) are often not included in these questionnaires (Moses, 2010).

Over the years, education statistics experts have developed, tested and implemented methodologies to deal with measurement error and missing data. Together, these methodologies provide well-defined steps to procuring reasonable-to-excellent estimates of the number of pupils who are attending schools in all countries (see section on data for coverage). The methods can be divided into two broad categories: first, steps to ensure the reliability of EMIS data directly; second, pupil estimates that rely at least in part on household surveys, as a complement to the EMIS data.

Methods to ensure reliability of EMIS pupil enrollment data.

There are several factors that can reduce the accuracy of pupil enrollment data, but effective EMIS systems can take measures to mitigate these sources of error (Moses, 2010).

1. Incomplete coverage of schools (namely, not all schools report data).

   - Based on the percentage of schools reporting, either the previous year's data for a school is inserted (since, if the school has not closed, its enrollment the prior year is likely to be close to that for the current year) or a percentage adjustment is made for unreported schools.;

2. Incomplete or inaccurate reporting (headmasters count incorrectly or falsify enrollments for financial reasons).

   - Inaccurate reporting is countered through a verification process in which, using a 2-5% sample, actual base records are checked by headquarters or regional personnel. This provides both an independent assessment, and a reference to original documents.

3. Declining enrollment over the year (enrollment is typically counted at the beginning of the school year, but enrollments often decline over the year.)

   - Enrollment is requested more than once—typically at each term, and some countries track actual attendance reported weekly or monthly, and make adjustments of pupil driven expenditures according to actual attendance.

4. Certain sectors of the system not counted in any annual census (such as adult education.)

   - EMIS analysts determine that all education sub-sectors are reporting, and ensure consistent "composition" of reported education sub-sectors.

This approach to data accuracy refinement is supported by recent work conducted in seven African Countries (UNESCO, 2010a).

Countries participating in the ICP could, at a minimum, be asked to document which of these four techniques were in use for which years. Based on this information, a decision can be made whether to proceed with a more in-depth data verification, and whether or not to offer support in providing resources and building institutional capacity in the country required to carry out data verification. An independent sample should be taken locally in order to judge accuracy (measure 2).

In a pilot stage for testing an output approach to estimating education PPPs, it may be useful to identify a few countries with concerns over the quality of pupil enrollment data, and to test the methods above for increasing data accuracy. This pilot will also provide an opportunity to assess the feasibility of country ownership, so that the data verification procedures could take place internally in the future.

Methods to estimate pupils where EMIS data are missing or deemed unreliable

Alternatively to EMIS, or as a complement to verifying EMIS data, the enrollment figures can be corroborated with data from other sources. In the absence of a national verification process or access to local information that would allow users to do their own verification, administrative data can be corroborated against household survey data and population censuses (the former are available for many more years than the latter), using an approach described below. Household surveys capture a wealth of information on expenditure, earnings, health, labor, agricultural output, as well as school attendance rates. Two international household survey series have earned a reputation for accurate data collection, and are a valid alternative source of information on education attendance rates: the Demographic and Health Surveys (DHS) funded by USAID; and the Multiple Indicator Cluster Surveys of child welfare administered (MICS) by UNICEF. The techniques to evaluate and improve the reliability of pupil data from household surveys have been primarily developed by UNESCO Institute of Statistics (UIS) and UNICEF (UNESCO and UNICEF, 2005; UNESCO, 2010b).

The UIS has a long tradition of comparing household survey attendance rates to EMIS enrollment rates. The EMIS enrollment *rates* are counted as: the number of pupils in a given level divided by the children of the official age for that level. The household survey attendance rates are: the number of children in the survey who were attending school[2] at a given level, divided by the total number of children in the survey corresponding to the official school age cohort. Discrepancies between the two numbers can arise from error associated with enrollment fluctuations, inaccurate attendance collection, incomplete inclusion of schools or population groups (e.g. complementary school models, homeless children); or incomplete or flawed population data that underlie enrollment rates. In addition, the fact that enrollment is not exactly the same thing as attendance may also affect the gap between EMIS and household survey data.

For their 2005 estimate of the number of primary age children not in school, UIS and UNICEF jointly developed a process for corroborating EMIS data with household survey data. First, the UNESCO/UNICEF method includes consistent definitions of school levels and school participation (section 7 below deals with school levels). If school levels and defined participation are consistent and if measurement error is minimal, then school attendance rates from a household survey and school enrollment rates from EMIS data for a given country, school level and year should be very close. UIS and UNICEF consider differences smaller than 5 percentage points to be acceptable, and see such small deviations as an indication that both sources of information are reliable. In such cases, EMIS counts of pupils can be used with reasonable confidence. For cases where a sizeable difference remains (>5 %) it is likely that one or the other data source has an error or an omission that needs to be reconciled. The criteria and a process to locate errors and decide which of the two sources was more credible were refined in UNESCO (2010b).

[2] Usually defined as children who attended school at least once during the year; in some surveys limited to children who attended school in the last week.

Many of the proposed improvements lie outside the scope of the ICP and are addressed to the bodies that collect data, but the process for deciding whether to use household survey data or EMIS data is relevant when the discrepancies between the two sources are substantial. UNESCO proposes that in those cases, local or regional education experts analyze the data and propose on a case to case basis which source appears more reliable. This paper would add to that proposal: if the EMIS data have already been validated with a method described in the previous sub-section, then the EMIS data should be given more weight in the expert consideration.

In some instances, EMIS data are completely non-existent or where they are enormously out of date, and household surveys become the primary sources of enrollment information. For example, among 47 low-income countries, two do not provide any enrollment numbers from EMIS systems – Haiti and Somalia[3]. However, these countries both have had a recent DHS or MICS survey that can be used to estimate attendance, and hence, derive an estimate of pupil counts (multiplying the attendance rates by the total number of children in the age cohort).

By using a combination of validated EMIS data and validated household survey data, a near-complete coverage of all ICP countries should be achievable for pupil counts, by education level.

3. Measuring the Quality of Education Services – how much pupils have learned

The second basic component of the value of education services is a measure of quality. The quality of education received by each student should be reflected in the output -- a rise in the level of skills and competencies acquired as a result of completing a course of study. As discussed in Pigozzi et.al. (2009), these skills and competencies are more than the acquisition of academic competencies in reading, mathematics, science and so forth. Pigozzi et al. state that "quality education needs to be viewed more broadly to include the range of purposes that are expected of education, including 21st Century skills such as team work, communication, negotiation, and foreign language ability, and societal priorities such as good citizenship, flexibility, and employability". Furthermore, equity (equal access to schooling regardless of race, gender, income or other factor) may need to be considered as a quality of the education system.

There are many proposals to measure school quality. Authors who have considered measuring school output as indicated by what pupils have learned include: Konijn and Gallais, 2006; Schreyer, 2009; OECD, 2007; Eurostat, 2001; Crouch and Fasih, 2004; Crouch and Vinjevold, 2006; Mingat et.al. 2004. But measuring output (learning) is not necessarily the only way to measure quality. Some authors propose using inputs to schooling as proxies to measure quality (e.g. Hill, 1975: Sergeev, 2007). International research on school effectiveness has demonstrated that inputs such as location of school, in-service training, the

[3] This number is from the EPDC/GMR Background paper for education projections made for the UNESCO 2010 Education For All Global Monitoring Report (EPDC and GMR, forthcoming).

provision of teaching and learning materials, and student assessment increase student learning outcomes (Lockheed and Levin, 1991; Lockheed and Verspoor, 1991; Pennycuick, 1998; Fuller and Clarke, 1994; and Hanushek, 1995). Other authors suggest using inspections or observations to measure school quality (Lequiller, 2005; Eurostant, 2001; Pritchard, 2002; Moore et.al. 2010). Another set of proposals focuses on progression and graduation rates (e.g. Fraumeni et.al. 2008; Atkinson, 2005). Some studies have also focused on earning comparisons (e.g. Fraumeni et.al. 2008; OECD, 200; Murray, 2007; Hanuschek and Woessmann, 2009). The earnings approach is intuitively appealing because it measures how much education skills are valued in the economy. Although there is considerable research in to effects of education on economic development (Hanushek, 1996; Woessman, 2007; Hanushek and Kimko, 2000), there is less conclusive comparative work on the relationship of education to employment. Also, education may not be the primary determinant of future employment – there are many labor market influences not related to education. Knight and Sabot (1990) find in an international study of East Africa that personal contacts are more likely to have a positive employment result than educational credentials.

The above approaches are not mutually exclusive and indeed, many of them can be combined to create a more robust approach to measuring education quality. This is particularly true (and even necessary) for the exercise of measuring education quality for the ICP's developing and transitional countries because the data for any one approach are not available for the whole set of ICP countries.

Arguably, learning achievement is the more immediate measure of the quality of education. Learning achievement measures are used in Eurostat's new education PPP approach; and for global consistency it is sensible to include them in the ICP's work also. Learning achievement measures from internationally comparable assessments are available for a growing number of developing and transitional countries.

That said, there are limitations to these measures. One obvious limitation is that assessments measure only a portion of what is learned. Another limitation is a likely selection bias for international exams in countries where not all students reach the grade in which the tests are administered, and those who have dropped out can be presumed to have learned less (e.g. Kees, 2000 and Fuller, 1987). If learning measures are to be used as quality proxies, then such factors need to be accounted for. However, the greatest challenge in using learning achievement measures to compare quality across borders is that there isn't common learning assessment available for all ICP countries.

Given this situation, inputs from other quality measurements – school inputs, a sample of classroom observations, contextual information (GDP per capita, demographics) can be used to corroborate and complement learning scores. The inclusion of other measures can potentially offer a more nuanced insight into school quality where learning assessments exist, and these measures can be used as proxies of learning quality where learning assessments are absent. This paper proposes a methodology to develop a common learning achievement measure by using all of the available information.

Measuring quality through learning assessments

Ideally, to measure learning, one would need a common metric for every country across the globe – a measure like the comparable quality of rice varieties. There is a growing number of large-scale international assessments of student achievement, which have especially proliferated since the mid-1990's. The coverage of one international assessment series, PISA (15-year olds preparedness for the labor market) is now wide enough that the OECD will use it as the main measure of quality for OECD education PPPs: "In the 2011 ICP the OECD will make estimates of educational output based on students at various educational levels with some quality adjustment based on standardized tests" (Deaton and Heston, 2009:39).

The largest assessments – TIMSS (math and science in 4th and 8 grades), PIRLS (reading literacy in 4th grade), and PISA – have tended to have greater representation of developed countries than the developing world, although this trend is slowly changing: in the latest round of TIMSS, nearly half of the participants were either developing or transitional states. In addition, regional assessments, such as Latin America's LLECE, Africa's PASEC and SACMEQ[4], provide a gauge of quality across each group of countries, but with weak generalizability outside of the region.

Error! Reference source not found. shows an overview of country participation in international and regional learning assessments. Because all of the available cases will be used to improve statistical power in producing a common metric, it is useful to examine the availability of learning outcomes data across a larger sample. This figure covers all 183 countries participating in ICP in 2011, including ones that are administered by Eurostat. The OECD countries in each region are shown separately from the non-OECD high-income countries, and the developing and transitional countries. In total, out of 183 ICP countries, including OECD, 90 have participated in one or more of the international assessment series (TIMSS, PIRLS, or PISA), and 109 participated in either an international or a regional assessment. Further, out of the 135 developing and transitional ICP countries, a total of 47 have participated in one or more internationally comparable assessment, and an additional 19 participated in regional assessments (Figure 2).

[4] LLECE is the Latin American Laboratory for the Assessment of Quality in Education; PASEC is the "Programme d'Analyse des Systèmes Educatifs des Etats et gouvernements membres de la CONFEMEN" and SACMEQ is the Southern and Eastern Africa Consortium for Monitoring Education Quality.

Figure 1. Participation in international achievement studies

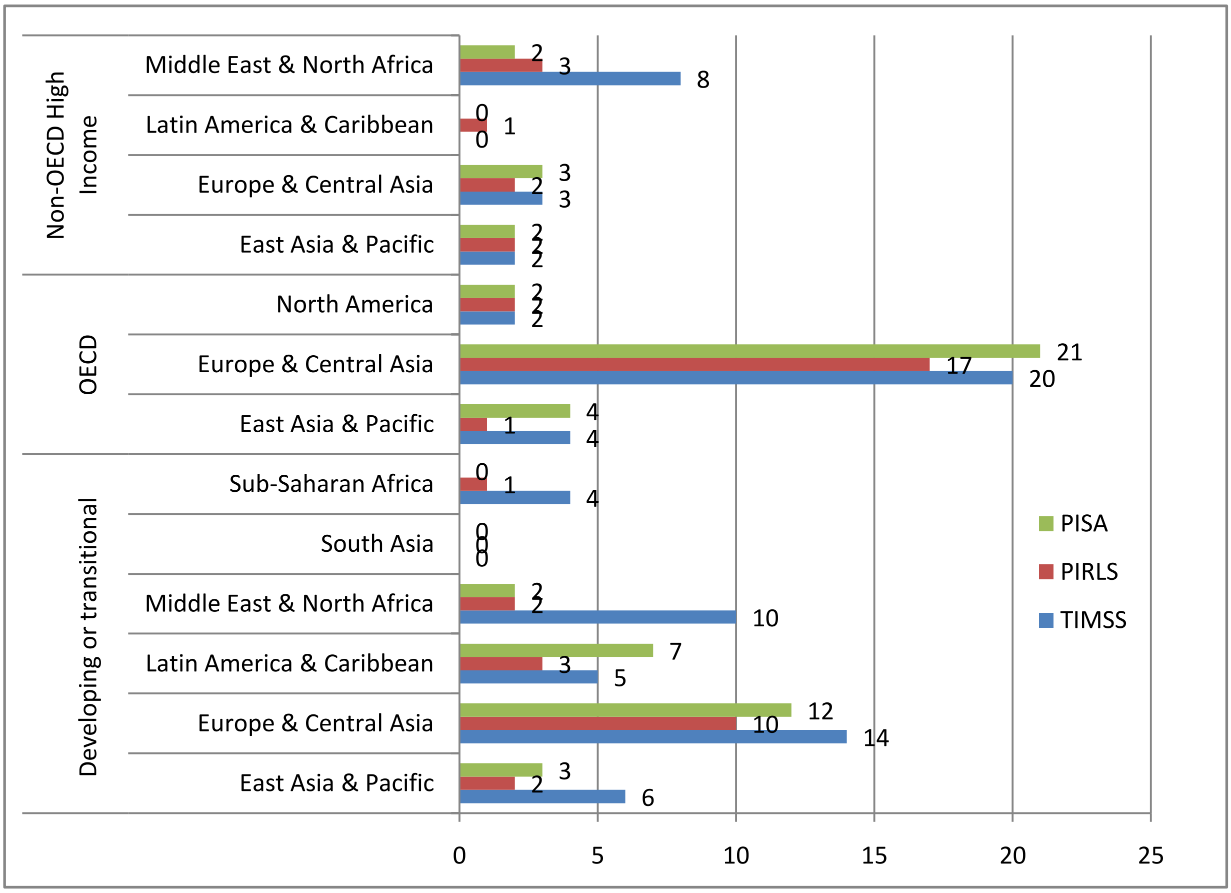
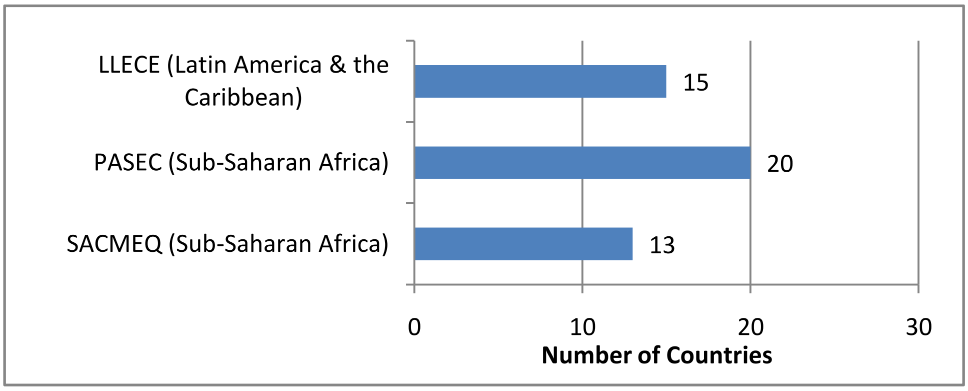


Figure 2. Participation in regional assessments



In conclusion, one common measure of learning does not exist for ICP countries, but it can be derived from the available information. The methodology proposed in this paper is intended to obtain such a metric for non-OECD (Eurostat) countries, but it will use all of the available data from the entire 183 ICP sample. The remainder of this section reviews two methods that have been applied to international achievement scores to derive a common learning measure for all countries included in the ICP: one proposed by Crouch (Crouch and Vinjevold, 2006), and by Hanuschek and Woessmann (2009). Both of these methods devise a common score from countries that have participated in *different* assessments. Neither addresses the case of 74 countries with no assessments.

Proposed methodology for the imputation of learning scores and development of a common metric

While the literature on missing data imputation is extensive, the use of imputation in development context has generally been limited, with studies using only cases with complete data to estimate relationships between variables of interest. However, a shared understanding of the bias introduced by ignoring missing values, particularly in studies where nation-states are the units of analysis, has driven researchers to consider data imputation as a way of increasing the precision of estimates and the generalizability of conclusions drawn from statistical inferences. This section of the paper discusses two of the methods that were employed in the literature to address the missing data problem in learning outcomes, and proposes a third, more robust strategy for handling missing learning scores.

Imputing common learning scores from different learning assessments - Crouch

Crouch (Crouch & Fasih, 2004; Crouch & Vinjevold, 2006) has proposed and tested a method for imputing learning scores over a cross section of countries based on both the international and the regional achievement studies. The studies include PISA, PIRLS, TIMSS, as well as regional assessments, such as SACMEQ (Southern Africa), PASEC (French-speaking Africa), and LLECE (Latin America). Crouch imputes an equivalent of TIMSS (the 1999 series) scores using a recursive regression method, where the TIMSS 1999 scores for countries that did not participate in that assessment were predicted from other achievement studies. The method proceeds in several steps, selecting assessments one-by-one, based on their degree of correlation with TIMSS 1999. The TIMSS 1999 scale was chosen as the target variable for imputation due to the greater representation of developing countries in this assessment (38 countries), as well as a relatively large overall sample of countries. Imputation of missing values was done using univariate linear regressions of the target variable on other assessments in a staged approach, gradually increasing the number of predicted values.

In the first stage of the imputation method, Crouch predicts the scores for countries that did not participate in TIMSS 1999, but participated in one of the other large-scale assessments, such as the IEA Literacy Study (1996) and the first rounds of PISA (2000), PIRLS (2001), and TIMSS (1995). To do this, the outcome variable, TIMSS 1999, was regressed on each of the assessments to generate predicted values. For countries with more than one predicted value generated (due to participation in more than one assessment), a weighted average of several predicted values is used as the imputed TIMSS 1999 score, with weights computed from the number of observations the assessment had in common with TIMSS 1999, multiplied by its correlation coefficient with TIMSS 1999. At this stage, the number of countries with actual or imputed values for TIMSS 1999 had increased from 38 to 64.

In the next stage, this larger number of outcome values is regressed on regional assessments with a lower initial correlation with TIMSS 1999, one assessment at a time, in another series of univariate regressions. Predicted values for TIMSS 1999 from these regressions either replace missing values, or are used to estimate a weighted-average imputed value, if a predicted value from the previous stage of imputation was already available. Ultimately, the TIMSS 1999 distribution with both the actual and the imputed values grew to 99 countries, of which 59 developing countries. Robustness checks performed by Crouch involved randomizing the final re-scaled TIMSS 1999 score to generate 30 alternative scores for each country, which were averaged to produce an alternative ranking of countries. The alternative rankings were found to deviate from the estimated, rescaled TIMSS 1999 rankings by an average of 4.5 ranks, which was seen as proof of robustness of estimates.

If the Crouch method were to be applied to the learning assessment scores available at the end of 2010, then the number of developing and transitional countries (LI, LMI, and UMI in World Bank classification) that would be covered is 66 overall; with 18 in Latin America; 15 in Sub-Saharan Africa; 10 in the Middle-East and North Africa; 6 in East Asia and the Pacific; and 18 in Europe and Central Asia

Imputing a synthetic common learning score – Hanuschek and Woessman

Another approach to the problem of insufficient data on any one international assessment, albeit restricted to a smaller group of countries, is recalibrating existing scores into a new scale, rather than imputing missing scores. Hanushek and Woessmann (2009) developed such a scale in order to estimate the effects of cognitive skills on economic growth. Data on countries that participated in larger-scale international assessments, including IEA's First and Second Mathematics, Science and Reading Studies (6 studies 1964-1991), three rounds of TIMSS (1995-2003), PIRLS 2001, and two rounds of PISA (2000-2003) were included in this analysis. The common metric was developed by transforming country means on various international tests into a new measure, based on a calibration formula that uses U.S. performance on the U.S. National Assessment of Educational Progress (NAEP) assessment, administered by the Department of Education since 1969 as a reference point. Given the fact that the U.S. participated in the international assessments that the authors include in their pool, and the data from NAEP assessments of American students are available for the same years, Hanushek and Woessmann re-calibrated the performance of all other countries that participated in international assessments to a new scale, based on their performance relative to the United States. First, the performance differences between NAEP 1999 and earlier NAEP years were calculated and expressed in PISA 2000 standard deviations. These differences were later used as adjustment factors for transforming the scores of other countries to the unified scale. Secondly, a new achievement scale was built by adjusting each country score on each test to a mean and standard deviation across a small group of OECD countries that participated in multiple tests (OECD Standardization Group, consisting of thirteen countries). Finally, standardized scores underwent temporal adjustment, using the U.S. performance differences in each test-year relative to NAEP 1999, computed in the first step. Using this final scale in regressions with economic growth as a dependent variable, Hanushek and Woessmann (2009) were able to argue that education quality had a strong causal effect on growth over time.

Proposal for an updated and expanded imputation of learning outcomes: Multiple Imputation

Both the Crouch method, and the Hanushek & Woessmann method were useful for the purposes for which they were designed: in the first example, plotting the general trend of the relationship between quality and country characteristics; in the second, illustrating the overall temporal relationship between educational quality and economic growth across a group of countries. However, these methods are highly deterministic, based on one variable at a time (one variable is used to determine the other), do not capture the uncertainty associated with imputation, and hence, may lead to erroneous conclusions regarding the properties of variables with imputed data. Furthermore, they cannot be applied to impute a learning measure for the 74 ICP countries that did not participate in an international learning assessment.

These weaknesses can be addressed with our proposed methodology, multiple imputation (MI). Multiple imputation holds invaluable advantages over other methods for handling missing data, as it accounts for the uncertainty around the values generated through imputation, allows for the imputation of several variables with missing values at a time, and provides flexibility in imputing variables that cannot be assumed to have a normal distribution (Rubin, 1996; Allison, 2002; Schafer & Graham, 2002). Furthermore, MI can accommodate separate multivariate conditional models for each variable with missing data. In the context of imputing learning scores at the country level, the ability to use multiple variables to predict the quality of education is essential, particularly when data from international assessments are not available for 74 of the total 183 countries included in the ICP 2011 process. With MI, variables such as GDP per capita or spending per pupil can and should be used to impute missing learning scores, alongside data from any direct assessments of quality. Other variables to include in a prediction model could be average years of education in the adult population, numbers of trained teachers, levels of teacher compensation (adjusted for PPP), share of youth in the age structure, and other variables presumed to be predictive of educational quality. The flexibility of MI relaxes the need for strict modeling and parsimony: the variables that are not predictive of the outcome would simply not affect the imputation, with no harm done to the estimates.

Unlike regression imputation, which is confined to available cases on the variables included in the imputation model, MI makes use of all of the available information in a given dataset, and sets up prediction models simultaneously for each of the variables with missing data. Imputations take place in an iterative process similar to a Markov chain Monte Carlo (MCMC) algorithm, where MI introduces a degree of uncertainty about the imputed values by generating several complete datasets. Imputation iterations continue infinitely until the distributions of imputed values (their means, variances, and covariances) converge, which increases the reliability of estimates and the validity of conclusions. The means of these plausible values are later used as the best estimates for the missing values of interest.

The major drawback of MI is its high demands on technology and time to produce reliable estimates, particularly in models with many variables and high proportions of missing data. Depending on available technology, MCMC chains may need to run uninterrupted for several days before convergence across imputed values can be achieved. However, once completed, the imputed datasets can be used for a variety of analyses, with results generalizable beyond countries traditionally participating in large-scale student assessments. While MI methodology, to our knowledge, has not been applied to development data, it is successfully used for addressing missing data in large-scale surveys administered by the U.S. Government (e.g. National Health and Nutrition Evaluation Survey (NHANES); Department of Transportation Fatal Accident Report System (FARS)). In sum, the properties of MI as an instrument for data imputation make it a worthwhile option to pursue within the framework of the ICP project.

Furthermore, many of the expected predictive indicators – which include GDP per capita, age structure, socio-economic status of families, pupil expenditures, pupil teacher ratios – are available for most of the ICP countries. Some of the expected predictive indicators, such as time on task, will require additional data collection. Whether the collection of those additional predictive indicators is essential and feasible is something that should be determined in a pilot study.

In summary, MI is a viable option, despite the high proportion of missing information. Achieving perfect convergence across multiple chains may be difficult, given the nonrandom nature of missing data (some countries are more likely to have missing values than others). However, even if perfect convergence is not achieved across multiple estimates, the ability to account for uncertainty, rather than rely on a single estimate, as well as the ability to retain the entire dataset by imputing the predictor variables would still increase the validity of the ultimate measure of the value of education. Single imputation (using regression) remains an alternative, but MI is preferred as a more robust method. The entire ICP/Eurostat dataset of 183 countries may be used to increase the statistical power for producing imputed values, the quality measure proposed in this section is geared specifically towards the ICP-only group. However, the ultimate group of countries for which reliable estimates can be obtained will be determined after these models have been run, as it depends on the availability of data on key predictors.

Estimation of tertiary level quality from imputed learning scores and relative earnings

The imputed common learning measure described above would be based on assessments taken during primary or secondary school. It is reasonable to use them as a quality measure across the whole of primary and secondary school (even if only one score is imputed). They cannot however, be a proxy for the quality of tertiary education, which is often organized differently than basic education, and for which strong selection procedures exist in many countries. Countries may have world-class universities, while the primary schools, on average, are poor. This is the case, for example, in Brazil.

There are no international output assessments for Universities. There are ranking systems, but they do not include many (if any) developing countries. Some wide-scale tests are applied in the United States, but not elsewhere -- GREs, LSATs – and the sampling of participants in these is biased as they are actually screening devices for entry into higher levels of tertiary education.

A proposed method to deal with the absence of tertiary assessments is to combine information from the imputed learning scores (above) and the ratio of earnings for young workers with tertiary compared to those who have only the level of basic education commensurate with the imputed learning score. The imputed learning score for basic education can be used to set each country a place on an internationally comparable learning scale. The ratio of earnings for workers with tertiary education compared to workers with basic education indicates the value added of the university education relative to the international learning scale. Through a simple transformation, the university education can itself be placed on a new, tertiary international learning scale.

The precise articulation of the function will need to be tested in the next phase of the project, when the actual model is developed. Data on earnings by income-level group can be obtained from Employment Surveys collected by the ILO or by via Household Expenditure Surveys, which are both conducted in many developing countries on a regular basis. The most appropriate age-group is that of young workers, for whom schooling is still their most important skills-asset. An alternative would be to use the incomes of some appropriately determined lagged group. Some adjustments may need to be made for factors such as scarcity of tertiary education (tertiary enrollment), migration, and quality differences between universities (if these are extreme and not averaged out in national numbers).

4. Adjustments to quantity and quality – time spent learning

It is widely acknowledged that pupil counts by themselves do not necessarily reflect educational volume because they do not consider how many hours pupils come to school; nor how many hours of education they receive on the days that they do come to school. Yet, the productive exchange between teacher and student that is at the core of education service production (Hill, 1975; Schreyer and Lequiller, 2007). The OECD (2007) notes the importance of collecting pupil hours (and also repetition and dropout) by level of education or grade to calculate education volume.

One measure of time per pupil is the official hours of school per year, available for all school systems via UIS. A large body of literature suggests that the actual hours of time in the classroom in developing countries is far lower than the official hours of school per year (Hill, 1975; Atkinson, 2005; Schreyer, 2009; Konijn and Gallais, 2006; Lequiller, 2006; Abadzi 2007a, 2007b; OECD, 2007; Fraumeni, 2008; Moore et.al. 2010). Furthermore, the literature finds that even in the classroom, time is not always spent on learning tasks (time on task), sometimes because the materials necessary for teaching and learning are not available. For example, Moore, et. al find that in over 100 schools surveyed in four countries more than half of the school year was lost due to school closures, teacher absences, student absences, late starts, prolonged breaks and other reasons. Abadzi reports similar findings. This is concerning, because it turns out that opportunities to learn are an important predictor of how much children learn and thus, of school quality (Schuh Moore et.al., 2010; Abadzi, 2007a; Woessmann, 2005).

Some countries collect information on some of the time lost due to absenteeism in attendance records. More countries could do the same (suggested by e.g. Chaudary et.al. 2006; UNESCO, 2010b). Beyond absenteeism, Abadzi (2007b) suggests using instructional time surveys for time loss adjustments. Fraumeni et. al. (2008) suggests an aggregation of pupil hours to account for actual service delivery to reflect a time component of educational quantity. Although Fraumeni, et. al., (2008) do not recommend methods for collecting actual teacher/pupil internaction time, others assume this component can be created through using official school contact hours (e.g. Eurostat, 2001; OECD, 2007). Lequiller (2006) suggests gathering this information via attendance figures, while another approach is classroom observation (discussed below).

One might argue that the time spent learning, or the opportunities to learn, per pupil, are already *a component of education quality* and need not be collected separately for the purposes of the education PPPs. Gallais (2006) argues that how much pupils learn (as measured in assessments, discussed next), already includes the pupils-hours component. This paper argues that it is still important to consider opportunity to learn because it can be used as *a predictor of learning, where other measures are absent*, and to *corroborate measures of learning* where they are available.

Classroom observation to capture opportunities to learn

Opportunities to learn are a measure of the effective time spent in the classroom – the combination of resources, practices, and conditions available to provide all students with the opportunity to learn material in the national curriculum. To capture the opportunities to learn, Moore et.al, use a classroom observation method using several instruments to collect their data including "Concepts about Print (CAPs); Early Grade Reading Assessments (EGRA); Stallings Classroom observation protocols; school observations; and interviews with teachers and principals" (p1). They collect data on: % of days school is open; student and teacher attendance; % of the school day available for instruction; % of student time- on-task; % of students with a textbook; observed textbook use; % of time spent reading; class size; school support, and, as a measure of output, grade 3 reading fluency. The findings allow educators to diagnose where instructional time is lost and thus, where improvements can be made. That said, time on task data does not present information on pedagogical performance; it simply provides information about how long students and teachers are on-task and what activities are happening in the classroom.

Research like this[5] is time and resource intensive. While in research studies, data can be collected from a small number of schools and data collectors can spend an entire day or two at the school collecting the data; at a national level, data collection would need to be done on a sample bases, requiring simplification of all instrumentation.

Such direct school observations may not be feasible for the ICP, but sampling of schools to collect some of the attendance and day-use measures may be feasible.

5. Adjustments for Non-school Factors that Contribute to Learning

Estimating the *value* of an education system, rather than its immediate output, requires controlling for outside factors that affect student achievement. It has been long established, for example, that the socio-economic status (SES) of a student is highly predictive of his/her learning outcomes (this vast literature dates back to Coleman, 1966). The simple logic is that families with educated parents and higher incomes would be more likely to support their children in learning, both in terms of helping them navigate the challenges of acquiring new knowledge, and providing the necessary physical conditions for children to be able to focus on schoolwork. For this reason, variables measuring student SES, such as various proxies for income level and parents' education, as well as immigrant status and ethnicity, are standard covariates in most individual-level production functions. In order for the education PPP to correctly reflect the value of education system irrespective of the characteristics of the student body, the country learning scores derived from student assessments must be adjusted for the effects of family support, or family capital. Adjusting for human capital will account for the large majority of non-school inputs to learning scores (it will indirectly also include such factors as child health, which itself is equally correlated with family capital).

In general terms, the imputed common score will be adjusted as follows:

$$Quality(1) = (\partial\ \times iScore) \times (\propto \times HC)$$

Where, the *iScore* is the imputed composite learning score, δ is a conversion factor for the *iScore* to Quality; HC is the average human capital of households or families with school-age children; and α is the adjustment factor for human capital, derived from the regression studies and MI.

In cross-country educational production functions with learning scores from international standardized assessments as the dependent variable, the effects of family characteristics are derived from student surveys administered in conjunction with such studies, and in the case of PIRLS, from parent surveys (see, for example, Fuchs and Woessmann, 2004; Ammermuller, Heijke &Woessmann, 2005; Schutz, Urpsrung, &Woessmann, 2008). The choice of measures used to proxy the effect of family characteristics in such analyses are restricted to the variables available in survey databases. In the case of TIMSS, for example,

[5] Conducted by organizations like the Academy for Educational Development (AED) and the Research Triangle Institute (RTI)

students are asked how many books they have in their home, what kind of items they possess among those defined as reflective of the socioeconomic status of the family (ranging from simple items such as a desk and dictionary to cars and student's own mobile phones), and so on. For example, books in the home were shown by Shutz, Urpsrung and Woessmann (2008) to be strongly predictive of student achievement in math and science.

Older students, such as the eighth grade population in TIMSS, as well as students in the PISA sample, are asked about the highest education level of their parents. Other studies include similar questionnaires gathering information about the families of the students within the sample. Shutz, Urpsrung and Woessmann (2008) argue that since in some countries, the levels of education do not fall neatly within ISCED levels, the reliability of parent education level as a predictor of educational achievement in cross country analyses is in question. Nonetheless, in many production functions using the achievement data, parent education showed strong independent effects on achievement over and above what is captured by books. For this reason, both of these measures may be combined into one family capital index.

In situations when data on these proxy variables are missing, which would be the case in all countries not included in the major student assessments, the family capital variables would have to come from other sources, such as Demographic Health Survey (DHS). Construction of indices based on available information from the DHS database would be reflective of the mean family SES in the same way as those constructed from student survey responses. The households to be included in the SES reconstruction will be those that have school-age children in school. The household capital index thus derived can also be calculated for countries that do have major student assessments, so that in the imputation process, the different predictive value of the household capital index from household surveys can be compared to the predictive value of the family capital index from the assessments. Combining both sources – using data from available achievement studies and DHS – would increase the validity of the ultimate "value of education" index, but achieving unbiased results using either of these approaches would likely to involve some data imputation.

6. Adjustments for Repetition, Dropout, and Length of Schooling

As mentioned in the discussion of time on task, the time that pupils have to learn is important. This is true within each year of enrollment, but also, the number of years that students are enrolled is a strong determinant of how much a student has learned. This rather obvious observation is relevant to the imputed common learning score, because countries vary in how many years the average student has been in school prior to reaching the grade to which the imputed common score applies. Say the imputed score applies to 8th grade, one of the grades tested in TIMSS. In countries with no schooling prior to first grade, the pupils will have had only 8 grades of school exposure; but in countries with one year of universal kindergarten pupils will have had 9 grades of exposure; in other countries where, on average, half of the children have had three years of preschool, pupils will have had 9.5 grades of exposure. The longer years of schooling should, if the quality of schooling is otherwise equal (and SES is equal), lead to a higher learning score. To obtain the correct estimate for average school quality, the learning score should be adjusted for the length of school exposure.

A similar logic applies to repetition rates. Students who repeat have taken (at least) two years to assimilate the material for one grade of school, but they are counted as two pupil-years. In the calculation of total education value, the learning obtained from each of the repeated years should be counted as only half. A simpler way to approach repetition is to count the learning in the repeated grade as zero. To obtain the correct values for the total value of school services, one approach is to simply subtract repeaters from the total pupil counts - the same mathematical result would be obtained if the iScores were adjusted downward by the repetition rate.

Putting both pre-school and repetition together to correct for the imputed score in general terms, will give (the actual equation may be refined and altered somewhat):

$$Quality\ (2) = Quality\ (1)\ \times \frac{1}{1+r} \times \frac{G}{AllS}$$

Where r is the average repetition rate up to the grade corresponding to the imputed learning assessment score; *G* is the corresponding grade (starting from grade 1 primary, and counting all grades up to the assessment regardless of the school organization, e.g. grades 4 and 8 for the TIMSS learning assessment); and *AllS* is the average number of years of schooling for students who reach the assessment including preschool and kindergarten but excluding repetition.

The above approach essentially assigns the same level of quality to each grade. It does not account for the bias – the fact that the students who reach the assessment grade may be those who have received better schooling, or have a higher aptitude than those who dropped out.

Another approach is to account for dropout in the quality part of the equation. Conceptually, this would presume that dropout is as much a result of poor quality as it is a result of family poverty, low rate of return on education, and other factors unrelated to what happens in the classroom. An education system that consistently fails a part of its population would be considered of lower quality, or lower value, even if reasonably good results are achieved with the remaining students. One reason is that due to the nonrandom nature of dropout, the students that remain in the classrooms are likely to differ from those who dropped out, and the bias created by this distortion will increase proportionately with the rate of dropout and repetition. Depending on the underlying assumptions about the causes of dropout, it may (or may not) be necessary to account for this bias on the quality side.

7. Stratification -- Accounting for Different Types and Delivery Modes of Education Services

In the previous PPP assessments, the ICP has stratified education as a whole according to school levels as defined in International Standard Classification of Education (ISCED) designed by UNESCO. ISCED is the internationally accepted norm for the stratification of education services. It divides education into the categories: Preschool (0), primary school (1), lower (2) and upper secondary (3), post-secondary non-tertiary (4), first level tertiary (5) and advanced tertiary (6). The current version used is ISCED 1997. The ISCED classification is well-developed, regularly maintained and acts as a standard around which sets of data from multiple levels and programs are organized. The different education levels provide comparable segments of the education system. Countries differ in the distribution of students across these seven levels. By appropriately weighting the number of pupils in each level, the overall education PPP can be estimated for each country. Moving forward to an output approach for valuing education services, the ICP should continue to use ISCED.

There are some possible exceptions to the use of ISCED – the system does not perfectly reflect all educational systems (Schneider and Müller, 2009). In a number of ICP countries, the grades or number of years included in preschool, primary or secondary differ according to the country's national system as compared to the ISCED definitions. In these cases, the decision of whether to use ISCED, may be more a matter of practical considerations. If the country's education data – including not just pupils, but also expenditures -- are only available along the national definition of school levels, it may be more expedient to use the national definitions. Most important to the PPP adjustments is that *all pupils benefiting from education services are included* in the stratification chosen.

The ICP has thus far accounted for public and private education separately, in part because different information for the input approach used presently is available for public education versus private. For the output approach, and the calculation of volume, quality, and the adjustment factors, the distinction between public and private may no longer be necessary. This would be the case if: a) aggregated pupil counts for public and private pupils are available; b) the imputed learning scores have been obtained from an appropriately mixed sample of public and private schools; c) the SES adjustments can be made for a weighted average of public and private school pupils; or d) the repetition, dropout and preschool enrollment rates are an average of public and private. If this were the case, then, even if there are distinctions between public and private schools, they will already be accounted for in the averages.

Besides the public and private distinction, it is important to ensure that other types of schools are included in the education counts, such as community schools (often funded by NGO's) or religious schools. Part-time and supplemental education services, such as music schools, sport schools, need to be included in the pupil counts, adjusted for their part-time status. If these are not collected by national systems, the ICP would need to include questions on these services in their country surveys. Alternatively, an estimate of the extent of such services for each of the ICP regions as a whole, supplied by local experts, could be added to the total education services to all countries within a region.

8. Overview of the Proposed Conceptual Model for Output Approach

The proposed conceptual model for an output-based educational value consists of two basic elements: 1) the volume, or quantity of educational services acquired; and 2) the quality of the output acquired as a result of these services, adjusted for the effects of non-education factors and the duration of schooling. At the highest level, the conceptual equation for the calculation of the PPP adjustment factor is:

$$Expenditures = PPPadjustment \times Quantity \times Quality$$

Where the Expenditures are known and collected through existing ICP procedures; the new methodology will provide Quantity and Quality measures, so that the PPP adjustment can be calculated.

Quantity

For the *Quantity* of education services the unit is pupil time in school, similar to the Eurostat output approach. This is intended to capture the actual amount of services received by the student population within a country, rather than the inputs (teachers, classrooms, supplies) invested in the system. Because not all pupils are full-time, the model will account for part-time pupils as a proportion of the full-time equivalents, based on the estimated hours of instruction relative to a full-time pupil[6]

$$Quantity = fulltime\ equivalent\ of\ Pupils = ftP$$

A possible refinement of the Quantity of pupils is an estimate of the actual time a pupil receives instruction, relative to say, some standard number of hours per year (this standard number of hours can be taken as the average of official school hours; or a benchmark for the desired number of hours). This refinement would account for time lost in formal schooling (which can be up to 50% or more of intended instructional time in some countries). The inclusion of school hours is a valuable component to understanding the components of education value; and may be a useful predictor for imputed learning scores. The refined equation is:

$$Quantity' = ftP \times \frac{actual\ annual\ instruction\ time}{benchmark\ annual\ instruction\ time}$$

That said, it may not be necessary for full-time formal schooling because, as mentioned in section 4, actual instruction time is also reflected in learning scores, which are used to proxy Quality (below). In other words, if instructional time, as a percentage of the benchmark school hours, is included in Quantity, then it must be controlled for in Quality – it is washed out.

Depending on the availability of data, the pupil count would be taken either from the official enrollment data, gathered through national systems such as EMIS and corroborated across several data sources to ensure reliability (see Section 2 above), or from official statistics of

[6] This is the ideal situation. In actuality, since information on part-time students is often not precise, the part-time student estimate may be an average based on a survey or desk-top expert study.

the school-age population, adjusted by the proportion of non-attendance, gathered from household surveys. The actual instructional time, or time-on-task, is a function of the official number of school days, less the number of days the school is closed during the academic year, and the amount of time devoted to instruction (official length of school day less the amount of time lost to non-instructional activities) and would need to be collected in sample surveys in most of the ICP countries, or estimated with an expert study.

Quality

The *quality* component of the model would be derived from three sources. First, the core element of quality is the actual or imputed composite learning score (iScore), which serves as a proxy of the education output. The iScore will be derived through a combined imputation method described in Section 3. It will be based on the latest rounds of the state-of-the-art international achievement studies, with actual values for countries where these assessments took place, and imputed values for countries that had not participated in these assessments, but for which the information on important predictor variables can be obtained. The predictor variables can be education input factors (e.g. pupil expenditures, pupil teacher ratios, books per pupil); or contextual variables found to have a strong correlation with mean learning scores (such as country wealth); or ones that are theorized to have such a relationship (e.g. time-on-task). The methodology for the imputation and the available data on learning achievement are detailed in Section 3 above.

$$iScore = f(learning\ assessments, input\ factors, contextual\ factors)$$

The iScore will be converted by an adjustment factor to a scale that is useful for the PPP adjustments:

$$Quality(0) = (\partial\ \times iScore)$$

In order to capture the portion of learning outcomes resulting from formal education, as opposed to the socio-economic characteristics of the student body, the iScore will be adjusted for a “family capital index”, which will include the mean level of education in the parent population (either on an ordinal scale such as ISCED, or in years of formal education), and mean family income level for families with children. The data for these adjustment factors can be obtained from student surveys administered as part of the large-scale achievement studies, or household surveys such as the Demographic Health Survey.

$$Quality(1) = (\partial\ \times iScore) \times (\propto \times HC)$$

Where, the *iScore* is the common imputed score, δ is a conversion factor for the *iScore* to Quality; HC is the average human capital of households or families with school- age children; and α is the adjustment factor for human capital, derived from the regression studies and MI.

In addition to the family human capital (*HC*) adjustment, the quality measure would be adjusted for the average years in formal schooling prior to 8th grade. The two key sources of variation in this element will be preschool -- in countries with large preschool enrollment, the average child will have taken longer to arrive at a given quality level – and the average repetition rate for the formal grades up to the grade of the learning assessment.

$$Quality\ (2) = Quality\ (1)\ \times \frac{1}{1+r} \times \frac{G}{AllS}$$

Possibly, dropout will also enter the equation as components of the quality measure, to reflect the bias in the socio-economic composition of the pupil body. This is necessary due to the highly nonrandom nature of dropout, and resulting distortions in the types of students that survive through primary and secondary grades, as compared to students that leave school before completing a full course of study. In addition, the quality of instruction may be theorized as a factor affecting the students' decision to drop out. These considerations have not been treated yet in the proposed conceptual model, and are not judged to be a major component of it, but will be taken up during the model refinement stage.

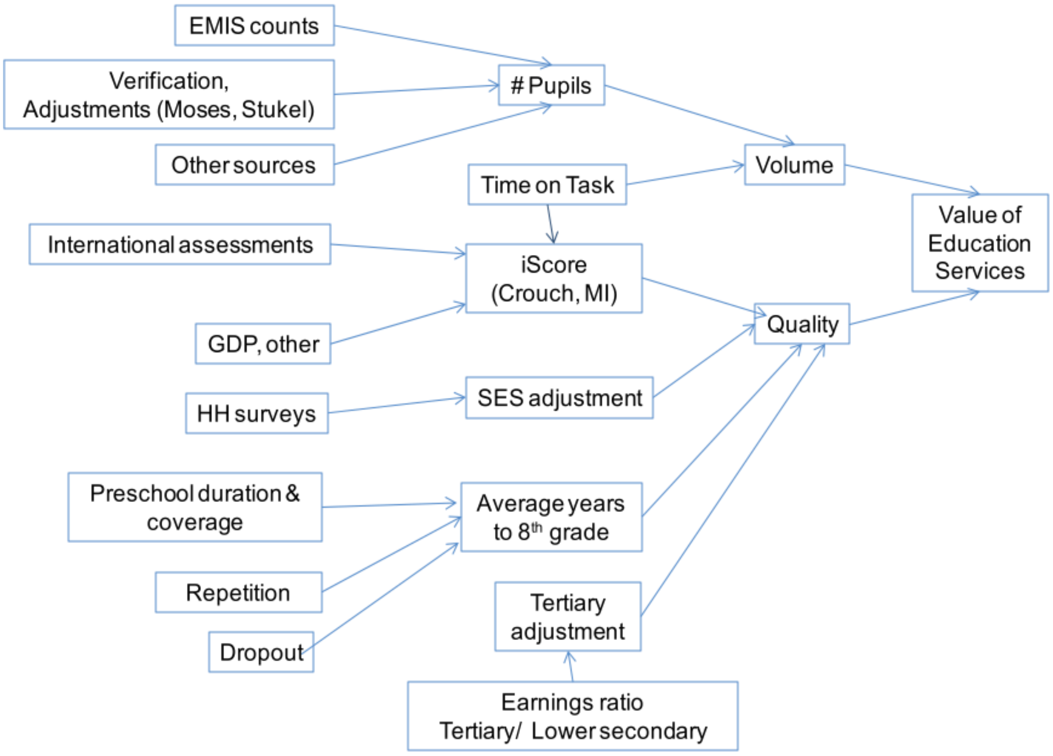
The third source of adjustment for the quality component is value added from tertiary education. Given the absence of comparable data on quality at this level of a national education system, and based on the assumption that tertiary education is strongly oriented towards the labor market, this element would be a function of the earnings ratio between individuals with and without a tertiary degree, and the scarcity of tertiary education in the country (i.e. tertiary enrollment). The earnings ratio would be confined to a younger age-group of workers (see Section 3 above).

$$Quality(T) = \gamma\ \times Quality\ (2)\ \times \frac{Earnings(T, 25to29)}{Earnings(LA, 25to29)\_}$$

Where *Quality(T) is* the imputed quality of the tertiary education system; *Earnings(T, 20-29)* are an average of earnings of workers with tertiary education, age 25-29, obtained from labor force or welfare surveys; *Earnings(LA,25-29)* are average earnings for workers with the education level equivalent to the grade of the imputed learning assessment; γ is an adjustment factor to bring Quality(T) to a scale appropriate for the PPP adjustment.

The whole model is shown again, in the form of a flow diagram in **Error! Reference source not found.**

Figure 3. Overview of conceptual model for output approach to measure the value of education serviced.


EMIS counts
Verification, Adjustments (Moses, Stukel)
Other sources
# Pupils
Time on Task
Volume
Value of Education Services
International assessments
iScore (Crouch, MI)
GDP, other
Quality
HH surveys
SES adjustment
Preschool duration & coverage
Average years to 8th grade
Repetition
Dropout
Tertiary adjustment
Earnings ratio Tertiary/ Lower secondary

9. Data Needs, Data Available, Local and External Expertise Required, Anticipated Coverage

Much of the data for the components of an output based approach to valuing education services are readily available, but some will need to be supplemented by expert imputation methods, collected locally through regional teams, or verified and adjusted by regional teams or international experts. This section provides an overview of the data needs, the availability, the gaps and how to fill those gaps.

Data needs

Data is needed to calculate the following components for the output approach: pupil enrollments at all levels; the imputed learning scores for basic education; imputed scores for tertiary education; socio-economic status; dropout, repetition and preschool; time on task.

For the **pupil enrollments,** EMIS data will be used for the majority of countries, supplemented by household survey data in cases where the EMIS data are not available, outdated, or unreliable. To ensure data reliability, local and regional teams will check whether the EMIS data have been verified. If no prior verification took place the teams will carry out the verification and validation of EMIS data in a sample of sites, using the techniques described in Section **Error! Reference source not found.**. EMIS data should also be checked against household survey estimates of school attendance, and where the two sources diverge, local experts will choose the source that is most likely to provide reliable information. Household survey estimates of pupils will be used if no recent EMIS data exists.

For the **iScore, available achievement data from** international learning assessments, as well as a set of contextual and input variables predictive of the quality of learning will be used to obtain imputed scores. The contextual and input variables include: GDP per capita and age structure - collected for all countries by UN agencies; socioeconomic status of the families (family capital) – available for most countries via household surveys, and for some, through learning achievement studies; pupil teacher ratios (PTR) are available for almost all countries from EMIS systems; per pupil expenditures available for all countries with national accounts and pupil estimates. One input variable that is important but not widely available is estimated time on task. To obtain this data, local teams would implement sample surveys in countries. The imputation process for obtaining the iScore will need to be done by an international expert.

The derived **Tertiary quality** will be obtained from the earnings ratios for young workers with tertiary vs basic education degrees, adjusted for the scarcity of tertiary education. The earnings ratios should be calculated with data obtained from labor or household expenditure surveys by local or regional experts. In those countries where the difference between public and private tertiary education is very large and is judged not to be captured in the average earnings ratios, a supplemental labor survey would be implemented among young workers with public and private tertiary degrees to obtain the earning differentials. An international expert should verify the calculations.

The **SES adjustment** for the iScore, to exclude private household contributions to learning from the education service valuation, requires data on socio-economic status of parents. These can be obtained directly from learning assessments, where those are available, or from household surveys to obtain a distribution of SES among parents. Near-full coverage of all ICP countries is anticipated. The adjustments for SES are well-documented in education literature and require intermediate-level statistical skills to implement by local or regional teams. An international expert can, if necessary, supply training and verification of the data used for the SES adjustment.

The adjustment for **years of schooling** prior to the learning assessment requires information on preschool attendance, grade repetition, and school dropout rates (as discussed in section 6). From these data, local and regional experts can calculate the average years of schooling pupils experienced prior to reaching the grade of the imputed iScore. Data on repetition and dropout rates are as commonly available as pupil enrollments, and near-complete coverage of all ICP countries is anticipated on these indicators. Preschool attendance rates, in contrast, are generally available for fewer countries than data on primary and secondary enrollment. In addition, preschool and early child care form a diverse group of services, ranging from private home care to full-fledged preschools integrated into the formal education system. Therefore, surveys administered by local and regional experts may be necessary to collect and standardize preschool attendance data for some ICP countries. The surveys may be designed with assistance from an international expert.

Data availability

As noted above, a large portion of the data for this exercise will come from secondary data sources, both at the country level and internationally. In sum, the following available data sources will be accessed :

1. International: large scale datasets (World Bank, UNESCO, EPDC), as well as international achievement studies

2. National: Education management information systems (EMIS), household surveys (DHS)

The data sources and required actions to increase the reliability and validity of the data are summarized in Table 1. The coverage rates shown here reflect the data available in the EPDC database; they are not exhaustive. In particular, it is anticipated that more sources will be found for the labor and welfare surveys necessary to calculate earnings ratios (Tertiary quality); as well as the household surveys for the SES adjustment.

Table 1. Data sources required actions to verify, adjust or impute data, and anticipated ICP country coverage.

Component	**Existing data sources**	**Supplemental sources**	**Verification, adjustments, imputations** **"Local" – task responsibility of local and regional experts.** **"External" – task for education expert team.**	**Present ICP Country Coverage**
Pupil enrollments	EMIS data, HH surveys	Verification samples and surveys	**Local:** check if data is verified; apply verification procedures; obtain surveys; analyze whether to use surveys or EMIS. **External:** verify procedures	96%
iScore	Learning assessments, GDP, SES, age structure, PTR, expenditure, time on task	None – data imputed	**Local:** use learning assessment score if available **External:** impute learning scores for core assessments; obtain iScore	100% goal; actual coverage TBD
Tertiary quality	iScore, labor surveys, HH surveys	Supplemental earning surveys for private vs public tertiary education	**Local:** calculateearning ratios; calculate iScoreTert with iScore, **External:** verify calculations	48%
SES adjustment	Household surveys, learning assessments	None	**Local:** Collect SES information; apply adjustment methodology **External:** verify adjustments	75%
Dropout, repetition, preschool	EMIS data, HH surveys	Verification samples and surveys, surveys for preschool	**Local:** check if data is verified; apply verification procedures; obtain surveys; analyze whether to use surveys or EMIS; calculate length of schooling **External:** verify calculations	90%
Pupil/teacher ratio	EMIS data	UNESCO	**Local:** use UNESCO pupil/teacher ratio data where available; calculate from EMIS otherwise. **External:** verify calculations from EMIS	Medium
Time on task	Attendance systems; studies	Sample classroom surveys	**Local:** implement classroom survey **External:** develop classroom survey with regional experts.	Low

10. Roadmap for implementing the pilot phase

This paper has presented a conceptual model and methodology for an output approach to measuring education PPPs in ICP countries. It is grounded in an in-depth review of the literature and methodologies to address similar challenges; was a collaborative effort that involved a team of education experts with complementary backgrounds; and included a review of the data as a first-pass assessment of the feasibility of the proposed approach. It is, nonetheless, a conceptual level proposal that will benefit from a critical review and guidance from members of the ICP team, the Regional Coordinators and the Technical Advisory Group, all of whom have longer experience than the EPDC in dealing with PPPs. Secondly, before the methodology can be implemented on a full-scale, it needs to be pilot-tested. This roadmap section briefly and broadly outlines the steps in this pilot phase. Should the ICP express interest in moving forward with the pilot phase, a more detailed proposal, level of effort and timeline will be developed.

Steps towards refinement of the conceptual model and methodology and pilot in small selection of countries

Below is a list of the steps to refine the conceptual model and pilot it in a small selection of countries. Some of these steps can be taken in parallel; also, some of the steps will need to be implemented by an external team of education experts; while others will be implemented by local and regional teams.

Steps 1-3 build upon each other and have a large external component to them.

1. (ICP/Regional and External) Refine proposed conceptual model of output methodology into an implementable instrument:
    a. Incorporation of feedback and inputs from ICP, regional coordinators, and TAG;
    b. Specify iScore imputation model and plan for implementation of methodology;
    c. Specify the general equations from Section 8, to include the complete set of indicators/data to be included;
    d. Develop methodology for verifying and/or estimating pupil counts.
    e. Develop methodology for survey to sample actual instruction time.
2. (External) First-level secondary data assessment and collection (for model-testing phase)
    a. Specific lists of data to be collected
    b. Compilation of data availability for all ICP countries.
3. (External) iScore imputation. Because of the nature of the process, the iScore imputation must be done simultaneously for all ICP countries after collection of the data.
    a. Test MI conditional models for subsequent imputation of learning scores.

b. Run imputations for first and second group of countries (see Section 3); and develop a composite learning score, the iScore;

c. Run imputation of composite learning scores for remaining countries.

Steps 5-8 can be implemented in parallel to steps 1-3 and have a large local/regional component.

4. (External) Assess whether there are important predictors for which data coverage is insufficient, and develop plan to collect this missing data. This step introduces some uncertainty into the process, as it is not known, at this time, what the most important predictors of learning scores will be for those countries where there are no learning assessments. It is not anticipated that there will be great surprises however, because the correlates to education achievements have, as mentioned, been studied in a large body of research, and the most important candidate-indicators will be included in the imputation process.

5. (Local/regional and external) Train local or regional technical teams to verify pupil counts, and to implement the sample survey for EMIS information verification.

6. (Local/regional and external) Train local or regional technical teams to assess actual instruction time, and to implement survey to collect missing information on instruction time.

7. (Local/regional) Assess whether EMIS counts or household survey estimates are the more reliable source of pupil count information; verify EMIS counts; implement sample survey to test EMIS information; adjust and/or calculate pupil counts by ISCED level.

8. (Local/regional) Implement assessment and surveys of actual instruction time .

Steps 9-10 can be implemented after steps 1-3 but in parallel to 4-8:

9. (External, but can train regional experts) Collect SES, repetition, dropout, and preschool attendance data for all ICP countries where such information is available.

10. (External, but can train regional experts) Calculate Quality and Quantity for all countries with complete dataset.

Steps 11-12 will be implemented after steps 4-10, and use information collected within each country.

11. (External) Rerun the imputation models with newly collected information, in particular, time-on-task.

12. (External) Evaluate the value added of time of task, and the difference from validated pupil enrollment numbers.

13. The final step will be to analyze and process the information into PPP adjustment values for the piloted countries, to evaluate the process, and to propose adjustments

where necessary for a full-scale implementation of the output approach to education PPPs.

The entire process will be documented for review and feedback.

APPENDIX – OVERVIEW OF INTERNATIONAL AND REGIONAL ASSESSMENT SERIES

- ### Progress in International Reading and Literacy Study - PIRLS

PIRLS is administered by the International TIMSS and PIRLS Center under the auspices of the International Association for Evaluation of Educational Achievement (IEA). This study tested 4th grade students for reading and literacy skills in 2001 and 2006. In 2006, 38 countries participated in the PIRLS and the next round is scheduled for 2011. A comprehensive survey of families was administered in both years as part of the study, providing a wealth of information not only on the student background, but also on the household practices contributing to reading and literacy

- ### Program for International Student Assessment - PISA

PISA tests science, math and reading skills of 15-year olds and has been administered four times since 2000, in three-year cycles (2000, 2003, 2006, 2009) . Organized and managed by the OECD, PISA is designed to capture and compare the preparedness of older students for entry into the labor force across countries. PISA has a diversity of academic content often not found in other international assessments and, similarly to other international studies, includes surveys of student and family demographic data which allows for the control of non-school factors during analysis.

- ### Trends in International Mathematics and Science Study - TIMSS

TIMSS is another assessment administered by the International TIMSS and PIRLS Center under the auspices of the IEA. Target groups are 4thand 8th grade students, and target areas are mathematics and the natural sciences

- ### Regional Assessments

The major regional assessments of learning outcomes are SAQMEC (Southern Africa), LLECE (Latin America) and PASEC (French-speaking Africa).