

HIPE-GMR Data Description

Version July 2010

	METADATA		PUPILS	2000	2001	2002	2003	2004	2005	2006	2007	2008	
Ethiopia 806	pre-primary duration	3	Males	1	1124386	1163731	1159879	1174839	1321818	1775832	1576988	1825817	
	primary duration	6		2	738052	798667	837995	832899	861051	1058586	1404533	1289111	
	lower secondary duration	4		3	586421	669264	720549	743631	732855	790817	946288	1201639	
	pre-primary entrance age	4		4	476600	550441	623188	675048	695416	707849	749826	865342	
	primary entrance age	7		5	372921	439490	502193	563922	602094	644605	671449	739684	
	upper secondary duration	2		6	256608	330846	385079	436582	470615	525937	564982	589010	
					7	214828	272196	353550	418977	460702	498029	532253	565974
					8	157454	191554	231502	297064	343570	389806	477576	508388
					9	118476	166593	209331	244458	292483	350404	405892	465647
					10	91849	73633	101791	127091	151278	203510	272242	295027
					11	81790	123097	24506	35861	35359	32120	57585	62401
					12	46412	67359	128703	20719	33830	35293	34107	54599
					13								
					14								
			Females	1	830649	903000	911553	937571	1132505	1591095	1396460	1607318	
				2	497771	586367	637477	640563	683119	907571	1266061	1144307	
				3	354470	445661	511377	543248	563016	633360	822344	1089867	
				4	265334	336274	411799	469506	510167	554177	616241	754168	
				5	203278	247370	300704	357371	415248	469874	522316	611905	
				6	140769	179730	211250	247894	291759	360026	434093	458951	
				7	132003	161616	201604	233055	271917	304828	361170	414993	
				8	110969	125266	144658	171095	196786	236249	309040	350742	
				9	92802	121844	133054	138052	158066	195382	233006	288563	
				10	64254	53376	68151	76708	84131	111438	154454	174425	
				11	49808	87105	12495	14484	11638	14353	18572	40453	
				12	26328	43167	86610	7991	13833	10717	13135	17766	
				13									
				14									
Gambia, The 813	pre-primary duration	4	Males	1	18954	16255	16973	17416			17187	17187	
	primary duration	6		2	15125	14253	14891	15278			15568	15568	
	lower secondary duration	3		3	14673	13933	14582	14960			14821	14821	
	pre-primary entrance age	3		4	13441	13119	13730	14085			15097	15097	
	primary entrance age	7		5	11565	12916	13560	13909			13208	13208	
	upper secondary duration	3		6	9909	11705	12325	12640			12303	12303	
					7	8108	8700	11275	10598	12196		10690	10690
					8	7229	7684	9962	9369	10781		11065	11065
					9	6957	7324	9520	8993	10349		11311	11311

Citation: Education Policy and Data Center, HIPE-GMR Data Description, 2010

ABOUT THE DATA USED IN HIPE-GMR MODEL

The most time-consuming and painstaking task of the costing exercise was finding the baseline data. The magnitude of the task was large – over 100 indicators were needed for 46 countries: a grand total of over 5,000 values for the baseline year alone. Ultimately many more data points were collected as the team also compiled values for multiple years – the model contains over 26,000 historical values. The data needed to be of the best possible quality and meet the tests of: accuracy, timeliness, transparency (easy to trace back to its source) and, in this study of 46 countries, international comparability. Five challenges required consideration in the data collection: data from multiple sources, resolving ambiguity, reconciling conflicting numbers, need for indirect calculation, and the need for estimation where no data could be found. The team resolved these challenges to the best of its ability, but recognizes that the data in the model are not perfect, and that even the baseline year values for the projections have some unknowable margin of error. That said, the team gave much thought to: the process of data collection; identifying and locating the (perceived) best numbers; cross-checking and triangulating data; and is convinced that, taken on the whole, the collected numbers are reasonable. Should readers find errors, please contact the GMR or the EPDC team; corrections can be made in an online version of the model.¹

The following pages explain in more detail how the team approached and resolved the data compilation task, and what data sources were used. An exact listing of the data sources is available in Appendix Table A of the GMR background report for this costing exercise, which can be found on GMR and EPDC websites.

Data from multiple sources. The numbers needed for the costing exercise were not available in one source, even for a single country. The team compiled them using many different resources including international databanks, national education and poverty planning or monitoring reports, special studies and reports (including some commissioned by the GMR), household surveys, estimates and expert judgment. Most of the resources used are available online; a few (yet unpublished) background papers for the EFA GMR 2010 report were made available to the team; and the GMR had special access to some national information in hard-copy reports.

Selection from multiple sources. For some indicators, more than one source was available, each with a different value; for the model one number needed to be selected. Although countries report numbers of pupils to the UIS, for example, there are some cases where the national education report has a different number.

In selecting from multiple data sources, the team was guided by an *order of preference* for different categories of sources. The most preferred sources were international databanks, because these data have been vetted by experts for international comparability². The second category of sources was national reports for education strategy and planning and national databases; the data in these sources are officially sanctioned for accuracy by the governments producing them. An issue with national data

¹ Contact information: epdc@aed.org.

² This does not guarantee accuracy; but raises the probability that the numbers are correct.

sources is that they are not necessarily internationally comparable as countries may use different definitions. The third category of sources was special studies and reports containing data on specific topics, which usually were themselves based on special access to national data sources. However, because the reports are one step removed from the original sources, national data was given preference, where it was available. In choosing between these first three sources, consideration was also given to *timeliness*, and in some cases, where a report or national source had more recent information than an international source, the former was used. Fourth, the team used household surveys to estimate pupils (via attendance rate and population numbers) in a few cases, but it is well-known that survey data often diverge by some percentage points from administrative data, so this source is used only as a last resort. Finally, when no information could be collected from any of these categories of sources, the team developed a few estimation procedures discussed below, to fill in the last remaining data gaps.

The most significant international source used in the model is the UNESCO Institute for Statistics (UIS). Data from UIS for all or most countries are used for the values of: school metadata such as official age of entry; duration of preschool, primary and secondary; number of pupils; percentage of private pupils; gross enrolment rates and intake rates; adult literacy rates; number of teachers and the pupil teacher ratio; and public education expenditure indicators. A second large international source is the United Nations Population Division, from which all of the data on population growth was taken, except for North and South Sudan (for these two regions, estimates from the UNDP provided by Chang et al. were used).

Resolving ambiguity. Sources are not always clear about their definitions and some of the indicators in the model are not necessarily clearly or uniformly defined— for this reason, international databases were preferred over other sources.

The indicator that is most vulnerable to ambiguity is government revenues – critical to determining the level of domestic resources available to education. A variety of sources gave values for government expenditures and government revenues as a percent of GDP that differed, on average, by a factor of two-thirds (where more than one number could be compared) and in two extreme cases by a factor of about four. Two reasons for these differences are principally (we believe) investments and donor contributions – both large portions of government education expenditure in some countries that may or may not be included in the government revenue or expenditure figure. It is also not always clear whether the government finance figure is for expenditure or for revenue. Because of the pivotal importance of domestic government finances to the financing gap, the team gave careful consideration to this indicator, and ultimately settled on an international source, UNSTATS, that has domestic government revenue data for most countries. The government finance baseline data has the greatest impact on the finance gap in the early years of the projections, when the projection values are closer to the starting values; for years closer to 2015, all government finance values are closer to the common targets for government expenditure set by the scenario.

Indirect calculation. Many indicators are not available directly, and have to be calculated using multiple numbers. For example, for many countries, the average teacher salary is not published, but it can be

calculated by dividing total numbers of teachers into total teacher salary expenditure; sometimes teacher salary expenditure had to be calculated from total education expenditure and the percentage going to salaries. The need for indirect calculation added to the number of indicators that were to be collected. The team used the same methods for the indirect calculations for every instance where they occur so the results are consistent across countries. A little over 300 values were calculated using indirect methods.

Unavailable indicators. Even using multiple sources and indirect calculation, some of the numbers needed were not available from official sources. For these numbers, the team had to develop credible estimates based on evidence. In some cases this meant taking a regional average and applying it to a few countries whose values were missing; in others, it meant making an estimate for all countries based on findings in the literature. Approximately 1000 values, out of 26,000 used in the model, are estimates. For some indicators, only a few countries were missing data and needed estimates: dropout rate, repetition rates, private enrolment, pupil teacher ratios, teacher salary, government revenues as a percent of GDP, public education expenditure as a percent of government revenues. For three indicators roughly three-quarters of the countries were missing data: pupil classroom ratios, non-salary recurrent costs, and distribution of public education expenditure by school level. For those indicators, where data were available for *other* countries, the team analyzed the data for patterns to define the estimation guidelines. For a few indicators almost no data were available and the estimates are based on special studies: classroom construction costs, additional costs to educate marginalized children, and costs of literacy programs.

Table B 1 shows the guidelines used to estimate the missing variables and the number of countries for which estimates needed to be made. The sources are divided into five main categories: international databanks, national reports and databases, special reports and studies, household surveys, and estimates and indirect calculations. Table B 2 shows the indicators relating to pupils and pupil flows. Table B 3 shows the indicators relating to teachers and classrooms and Table B 4 shows the indicators relating to unit costs and the budget. Within each table, the category that provided data for the majority of countries is highlighted in bold; whereas for categories that were used for only a few countries, the list of countries is shown.

Table B 1. Estimation guidelines used in the model for missing data.

Indicator (# countries missing)	Estimation guideline
Dropout rate (4)	Regional average used
Repetition rate (3)	Regional average used
% Private enrolment (13, 5) ^a	Regional average used
Pupil teacher ratio (6, 2) ^a	Regional average used
Government revenue (2)	Regional average used
Public education expenditure (3)	Global average used (no regional patterns observed)
Teacher salary lower secondary (17)	1.35 x primary teacher salary (global ratio)
Teacher salary preschool (41)	Equal to primary teacher salary (global ratio)
Non-salary recurrent costs per level (39, 29) ^a	For countries with non-salary recurrent costs all levels combined, use this combined value; for countries with <i>no</i> non-salary recurrent cost data (7) use regional average.
Public education expenditure by level	Expenditure for preschool based on ratio of preschool/primary pupils assuming equal expenditure per preschool and primary pupil; Expenditures for primary and lower secondary distributed according to duration of level and assuming expenditure twice as high in lower secondary as compared to primary following Lewin (2008).
Pupil classroom ratio (30)	Equal to pupil teacher ratio
Classroom construction costs (43)	Based on Theunynck (2009)
% Marginalized children (46)	Marginalized children assumed to be equal to percentage of marginally educated adults from the GMR deprivation and marginalization in education (DME) database.
Marginalized pupil costs (46)	Incremental costs based on: Chen and Mulkeen, 2008; Mulkeen, 2009a; height of direct subsidies to households based on Bundy et al (2009), Fiszbein et al (2009) and Chanamuto (2009).
Costs per literacy pupil	Based on Van Ravens and Aggio (2005, 2007)

a. First number for preschool level, second number for primary and lower secondary.

Table B 2. Sources for indicators related to pupils and pupil flows. Main source(s) indicated in bold; specific countries for values based on less common sources.

Indicator	International databanks	National reports and databases	Special reports and studies	Household surveys	Estimates and indirect
Population	UN	North/South Sudan only	1.	2.	3.
School metadata	UIS	4.	5.	6.	7.
Number of pupils	UIS	Rwanda, PNG	DRC, North & South Sudan, Nigeria	Haiti, Somalia	8.
Intake rates	UIS	PNG	DRC, North & South Sudan, Nigeria	Haiti, Somalia	9.
Repetition and dropout rates	Based on UIS	Gambia, Sierra Leone ^a , PNG,	DRC, North & South Sudan, Nigeria	Haiti, Kenya, Liberia, Rwanda, Uganda, Zimbabwe, Somalia	Afghanistan ^a , CAR ^b , DRC ^b , Guinea-Bissau
Transition rates	Based on UIS	Gambia, PNG, Rwanda,	DRC, North & South Sudan, Nigeria, Sierra Leone,	Haiti, Kenya, Liberia, Somalia	10.
% Private school enrollment	UIS	Tajikistan, Bangladesh, Ethiopia, Liberia, Zimbabwe	DRC, North & South Sudan, Nigeria, Haiti	11.	Afghanistan, Myanmar, PNG, Somalia, Uzbekistan
Literacy	UIS or CIA Factbook	12.	13.	14.	North & South Sudan
Adults in literacy training		Various national reports	15.	16.	Where missing assumed zero

Table B 3. Sources for indicators related to teachers, classrooms, and non-salary recurrent costs. Main source(s) indicated in bold; specific countries for values based on less common sources.

Indicator	International databanks	National reports and databases	Special reports and studies	Estimates and indirect
Pupil teacher ratio	UIS	Guinea, Malawi, Uganda, Haiti ^d , Ethiopia, Madagascar, Mozambique, Rwanda, Senegal, Tanzania, CAR, Vietnam ^c , Zambia ^c Zimbabwe	DRC, North & South Sudan ^d , Nigeria	PNG ^c , CAR ^c , Cote d'Ivoire, Haiti ^c , Malawi ^c , Mozambique ^c , Somalia, South Sudan ^c ,
Pupil classroom ratio		Rwanda, Togo, Burundi ^c , Vietnam, Sierra Leone, Mozambique, Bangladesh, Kyrgistan, Ethiopia, Senegal, Uganda, Yemen, Haiti, Zimbabwe	DRC, North & South Sudan ^d , Nigeria	Estimates based on PTR

Table B 4. Sources for indicators related to expenditures and budget. Main source(s) indicated in bold; specific countries for values based on less common sources.

Indicator	International databanks	National reports and databases	Special reports and studies	Estimates and indirect
Teacher salary – sources spread evenly	Pole de Dakar	Calculations based on national data or direct national data	Bennel (2009), Chang et.al. (2009), Mulkeen (2009)	Based on regional averages
Non-salary recurrent costs by school level		Bangladesh, Cambodia, CAR, Ethiopia, Mali, Mozambique, Nepal, PNG, Rwanda, Sierra Leone, Tajikistan, Togo, Zambia, Zimbabwe	CAR, Nigeria, North & South Sudan	Regional averages or average non-salary recurrent costs for all levels combined
Classroom construction costs by school level		Indirect: Burkina Faso, Chad, Haiti, Mozambique, Niger, Sierra Leone, Togo Direct: Ethiopia, Pakistan, Zimbabwe	CAR, Nigeria, North & South Sudan, PNG.	Estimates based on van Theunynck (2009)
Literacy training costs				Estimates based on Van Ravens and Aggio (2005, 2007)
GDP	UNSTATS	17.	18.	North & South Sudan
Government revenues as % of GDP	IMF, World Bank	19.	20.	North & South Sudan, Somalia
% public expenditure to education	UIS, Nationmaster, Pole de Dakar, World Bank	Vietnam, Zimbabwe	Martin & Kyrili	Afghanistan, PNG, Somalia
% education expenditure to preschool	UIS	Haiti, Sierra Leone	DRC, North & South Sudan, Nigeria	Estimates based on primary costs

% education expenditure to primary & lower secondary	UIS, World Bank^e	Haiti, Sierra Leone, Tanzania ^f ,	DRC, North & South Sudan, Nigeria	Estimates based on Lewin (2008)
---	------------------------------------	--	-----------------------------------	--

- a. Repetition only
- b. Dropout only
- c. Preschool only
- d. Primary and lower secondary only
- e. Pakistan only
- f. Primary only