**Audit Options to Certify Results for a "Cash on Delivery" Contract in the Education Sector**[1]

Luis Crouch
Jonathan Mitchell[2]
RTI International

Version of June 30, 2008
(Version 2)



## 1. Introduction

The Center for Global Development has proposed a "Cash on Delivery" (also known as Progress-Based Aid in some early discussions) approach, whereby, in a nutshell, a donor, or group of donors, would enter into a contract with a country under the terms of which the country would be rewarded for making progress on some internationally-agreed indicators based on an independently-audited statement by the recipient country. The education sector has been

[1] Submitted by RTI International staff members as a contribution to the Center for Global Development's effort to promote discussion of the notion of a Cash on Delivery approach to international development assistance.
[2] Research Vice President and Senior Economist, and Senior Education Researcher, respectively.

proposed as a possible "test case" for this approach, both as a thought experiment and possibly for a pilot of the approach if the concept gains traction. Early attention has focused on the number of primary completers and the number of students appearing in standardized primary school examinations as possible indicators of progress that would form the basis of the reward amount under the terms of the contract.

Some papers describing the Cash on Delivery approach call it a "hands off" approach to development aid, and it is worth stopping at this point, because much of what follows hinges on this aspect. The assumption here is that a great deal of donor and country effort is currently wasted with an approach that is too "hands on." Donors take a detailed interest in the minutae of how countries run their education systems, negotiating policy conditionalities (in a policy-based or SWAp mode), discussing project inputs and monitoring projects (in a traditional project mode) in sometimes excruciating and time-consuming detail. Not only does this consume time and effort, and distract recipient countries from getting on with their business, but it tends to assume that donors actually know a great deal about the "production function" of education (or any other development goal). Even if they actually do know, which may be subject to questioning, providing pre-specified conditions and project components tends to prevent countries from discovering and "feeling their *own* way" towards what works, particularly if many things work. And, it is likely that finding one's own way is a much better way to learn about what works. Thus, the "hands off" approach accepts that donors might not know that much about how to produce results, or that, even if they do, stimulating local discovery about "how to produce," and stimulating a multiplicity of efforts in how to combine inputs to produce results, is a good idea. This requires not spending too much time pre-specifying project inputs or policy conditionalities, and instead focusing on rewarding results: "hands off."

But because so much rides on the final result, the success of the Cash on Delivery approach is very dependent on the ability of an eligible country to gather reliable data on the agreed outcome indicators—data that could be independently verified within an agreed confidence interval. The verification mechanism, or audit, therefore plays a critical role in the Cash on Delivery approach, and the credibility of the audit will largely determine the credibility of the approach. The aim of the discussion below is to briefly clarify the purpose of the audit, examine the audit requirements of the indicators that are under discussion, identify some of the strengths and weaknesses of several audit options, and make some specific recommendations on the design and implementation of the audit based on operational feasibility and cost, to some extent. A guiding principle throughout the discussion is the hands-off approach that is at the heart of the Cash on Delivery approach, and we therefore steer away from audit requirements that become excessively interventionist or complex or that could divert attention away from the progress that is being measured.

Version 2 of this paper incorporates comments on Version 1, but may contain errors or inconsistencies.

## 2. Audit Purpose

At a very basic level the role of the audit is to convince the donor that the reported increments in progress by the country in question represent reality on the ground. The audit must therefore be able to verify the reported numbers (either by checking a random sample of the reported data or by generating independent estimates to which reported figures can be compared), confirm that they are based on a functioning information system that does not include fictitious students or ineligible examination takers, and certify that the increases in the numbers are not the result of cheapening the outcome or of gross gaming such as importing students from neighboring regions. Given that the progress to be rewarded must be measured against a baseline, the agreed baseline numbers must also be certified in some fashion even if not all audit requirements for future progress measurement can be applied to the baselines. If the baseline figures prove to be overestimates, the country may find that reaching a verifiable level of progress above them is too difficult to achieve. Conversely, baseline figures that represent underestimates would result in a windfall to the country, though (presumably) only in the first year and the windfall could be capped in the terms of the contract.

Within the Cash on Delivery framework, one requirement is that audited figures, and maybe the broader audit findings as well, be published locally to the greatest level of disaggregation possible. This aspect of the audit encourages local accountability, and some of the recommendations that follow, therefore, include audit elements in support of the social audit function even where not strictly required for verification purposes (and we identify them accordingly). In this way the audit can perform a social audit function as well, providing parents and civil society with the information necessary to raise issues of equity, quality, subtle forms of gaming or unintended consequences in ways that are as effective, as forces for reform, as the financial incentives from the donor. One could even argue that if the Cash on Delivery framework does not result in greater local accountability, it will probably not work in the long run in any case.

## 3. Audit Considerations

It is tempting to view the audit primarily as an assurance against outright misrepresentation or overt gaming by the eligible country, but an audit can serve a more fundamental development purpose as well. In many developing countries there are often very basic capacity and logistical constraints that may present significant obstacles to establishing baselines or measuring progress, independent of any deliberate misrepresentation. The South Africa example is instructive here. Despite considerable investment on data in the Education Management Information Systems (EMIS) that draw on two annual surveys of all schools, there is considerable debate regarding the

actual number of schools in South Africa, not to mention numbers of primary completers. Remarkably, the South Africa National Department of Education is still pondering apparent discrepancies between the figures in the1996 and 2001 "School Register of Needs" and also in the Annual School Survey of the last few years, whose resolution has continued to elude the country despite high level assistance by internationally recognized organizations such as PricewaterhouseCoopers. The number of schools in many countries is not known with absolute certainly, nor can many countries produce a list of school names. Likewise, both Nigeria and Pakistan have enjoyed considerable donor support for improvement of their education information systems, yet there remains considerable uncertainty regarding even very basic national education statistics in these countries. Honduras literally does not know how many teachers it pays. In such circumstances the donor may wish to utilize the audit function to both generate estimates of progress (that could even form the basis for payment in the early stages) as well as to examine and assist in the strengthening of standards and practices governing education record-keeping and information flow that would form the basis of future payments.

In the discussion of audit requirements for several indicators that follows, we have found it helpful to first identify likely capacity constraints and gaming opportunities (some of which are an outcome of capacity constraints) and then consider how an audit can address these. Based on the Cash on Delivery documentation available, the indicators we have considered are total primary school completers, total national standardized test-takers, and national standardized test scores. . For the reader's convenience we have summarized for each indicator in the table below a list of possible capacity constraints, likely gaming opportunities, and the resulting audit implications.

**Table of Audit Options for Cash of Delivery Contract in Education**

| Progress Indicator | Capacity Constraints | Gaming Opportunities | Audit Implications |
|---|---|---|---|
| **Number of Primary School Completers** | • No precise definition of completion<br>• No completion certificate<br>• No record keeping standards<br>• Inconsistent record keeping (as an example, South Africa does not even have an unambiguous master list of all *schools*)<br>• Low accountability<br>• Weak capacity makes it more difficult to control for perverse incentives<br>• Weak link between school and parents (parents lose or do not receive completion certificates) | • Cheapen completion definition<br>• Increase class size<br>• Lower teacher qualifications<br>• Further relax enforcement of, e.g., attendance as a completion requirement<br>• Lower promotion standards<br>• Open under-resourced schools<br>• Issue undeserved completion certificates, e.g., to children with minimal attendance and unmeasured | **Process verification**<br>• Have to audit to a formal definition of completion–impossible without precise definition. Unclear what happens when there is no formal definition.<br>• A completion definition will have record-keeping implications regarding attendance and promotion in the current and prior years. The audit would have to verify adherence to record keeping standards and confirm that records support claims of individual student completion.<br>• Verifying adherence to record keeping standards will require random checks of sample schools throughout the academic year. Audit implications of non-compliance are not easy to define. Examples in the discussion below illustrate complexity.<br>• Challenges of verifying attendance, or validity of process and records, may require more and more effort into something that largely certifies physical presence and which ultimately may not be that valuable.<br>• Establishing baselines will be difficult regardless of audit methodology.<br>• Establishing baselines could be explosive if current completion estimates are shown to be over-estimated or unsubstantiated by school records (e.g., even if there is a formal definition of completion that requires minimum attendance, *current* completion data are not justified by attendance records).<br>**Verification through HH survey –**<br>• Relevant households (i.e., households with children of school-completion age) a small subset of total – would require large sample size or creation of specialized sampling frameworks.<br>• Survey may not spot cheapening of definition, only gaming of the numbers.<br>• Difficulty resolving disagreement between official and household reporting may be insurmountable at anything like a reasonable cost. |

| Progress Indicator | • Capacity Constraints | • Gaming Opportunities | • Audit Implications |
|---|---|---|---|
| **Number of Students Taking National Standardized Test** | • Logistics<br>• Lack of student identifier<br>• Lack of control over who appears in examination (children or even older graduates taking test not in grade)<br>• Difficulty aggregating numbers | • Allow adults to take exam<br>• Allow older children to take exam<br>• Have younger children take exam<br>• Have out of school children take exam<br>• Double count children<br>• Import children from other regions or schools | • A unique student identifier (national ID number) matched electronically with test-takers would eliminate many gaming opportunities.<br>• Student ID would prevent student taking test twice (either in current year, or taking it in future years).<br>• With student ID problem of overage children taking test would disappear over time.<br>• Conduct retest of random sample of schools shortly after official test to:<br>➢ Compare enrolment with count of test-takers and count of retest takers<br>➢ Compare test-taker identity<br>➢ Control for grossly overage (adult) test-takers<br>➢ Control for ghost test-takers (say by attributing tests to out-of-school children or children otherwise not present)<br>➢ Verify status of claimed test-takers not present at retest<br>➢ Compare test scores with retest scores to discourage gaming such as testing out-of-school children<br>➢ Retest sample schools have to be truly random, or the system could selectively fake tests from schools not deemed likely to be retested.<br>• Require retest scores to be published to enhance social audit function.<br>• Could report names of schools whose retest-taker count is below the claimed test-taker count by x% |
| **Scores on National Standardized Test** | • Test security<br>• Test development to ensure comparability over time | • Cheating<br>• Teaching to the test<br>• Selective presence<br>• Lowering cognitive load of test over time | • Auditing test scores discourages gaming of test-taker numbers by including non-school goers or over-age people who wouldn't perform well on the test.<br>• Retesting controls for cheating and selective presence.<br>• Teaching to the test is not subject to mechanical audit, particularly since some of it is desired and it is difficult to say, in advance, how much of it is desired.<br>• Report test-taking numbers and penultimate year enrolment to discourage inter-year selective presence through forced repetition or "push-out." The incentive for inter-year selective presence can be balanced by the reward for test taking.<br>• Ensure that the incentives for test taking make it to the school. |

## 4. Audit Options

### 4.1. Primary School Completers

#### 4.1.1. Definitional Issues

Given the importance attached to primary school completion in the MDGs and EFA declarations and the widespread use of national statistics that purport to measure it, primary school completion is an appealing indicator for consideration for a Cash on Delivery contract. Upon closer examination, however, we note that there is often considerable imprecision in the definition of primary school completion, and even where definitions are precise, there are difficult audit issues associated with this indicator that are not easily addressed.

In a very general sense, primary school completion is understood to mean that a child has completed the "normal" primary school program of a given country that typically involves a progression through the mandated five or six (or up to 8 in some countries) primary school grades. In addition, there may be a requirement that the primary school or teacher in some way vouch that that the completer has mastered some basic skills or body of knowledge. This mastery requirement may simply be based on the teacher's opinion, or it may include a formal written assessment or examination (which may be school-specific or a national examination). Primary completion may also result in a formal certificate for the child (based on a score on a national exam or simply based on teacher or school opinion), certifying that he or she has met all primary school completion requirements. Such certificates are often required for a child to seek admission to the next tier of education. In addition to formal primary schools, many countries recognize non-formal education programs as equivalent to a primary education, provided that the programs maintain certain standards that often include a student assessment. In the discussion below we examine the audit related issues surrounding completion based on attendance only, and address the topic of testing separately.

To better understand the issues, it is useful to distinguish two extreme cases. One, where a country in fact has no real "certification" as such, or only some form of certification based on extremely informal information such as teacher opinion. In this first case the opportunities for gaming are rife, and it is actually difficult if not impossible to imagine what an audit of the number of completers would even mean. Only audits against standards make sense, and if there are no standards there is really nothing to audit. In these cases, the country could just, in essence, declare more completers, and since there are no standards, there is no way to judge whether these completions are in any way meaningful. The meaning of Cash on Delivery in these countries would be hard to discern, and maybe the Cash on Delivery experiment would simply have to be careful not to work in such countries. In the second case there is some formal definition of completion and at least a putative means of certification that a child has met the requirements for completion. In this case, in order to verify that a given child is a primary school

completer, an audit needs to verify that the student has met the defined set of agreed primary school completion requirements (which could be specified in the contract – as a spur to the development of standards, which could be a benefit of Cash on Delivery – or could be accepted on the basis of written national norms). At a very basic level, the student must have met the requirements for promotion through the sequence of primary grades (to avoid having children skip grades without justification or against norms, so as to pump up the numbers of completers in the short run), which, at a minimum, must specify a minimum number of days of attendance (to prevent promotion of students who in practice have barely attended) even if there are no learning requirements (as in the case of automatic promotion policies that prevail in many countries). Most countries have a mandated number of official schools days, and promotion (particularly automatic promotion) is meaningless if there is not a minimum attendance requirement. If there is no official minimum attendance requirement, the contract will need to specify one to prevent a country claiming completion for children who may have enrolled but literally or almost literally never attended.[3] At an extreme, these countries devolve to the no-formal-requirements situation of case 1, and in these cases Cash on Delivery (or audited Cash on Delivery) based on completion (but not based on test-taking, as explained below) may be a questionable notion.

### 4.1.2. School Records-Based Audit

In order to verify that the child has met promotion requirements, an audit must examine school records (to include at least attendance, but also any other official requirements) and confirm that school records are being kept according to standards, that these records are being applied to promotion decisions, and that government figures are consistent with school records. This confirmation would require random unannounced spot checks of schools over the course of the academic year, not simply at one point in time. The audit cannot simply confirm the existence of completed attendance records as these are extremely easy to manipulate in any number of ways. In order for records and record-keeping systems to be auditable, there must be standards against which they are audited. For example, and this is one example only, one such standard would need to specify the time of day by which the attendance record must be completed and in addition that attendance records must always be kept at school (or must always be at school by the time attendance is taken and be kept at school all day thereafter). To illustrate the importance

[3] Minimum attendance requirements should be treated with care. Serious investigation of student attendance as a requirement for promotion might lead to some embarrassing results. One might think that requiring attendance for 60% of mandated school days would be a conservative promotion requirement (120 days out of 200), but a close look at the number of days a school is closed due to bad weather, local holidays, elections, strikes, or classrooms closed due to teacher non-attendance, etc., may in many countries greatly reduce the number of days the school or classroom is even in session. If one adds to this the days that teachers are absent (and should these count towards promotion requirements?) it may be necessary to disqualify entire schools even with conservative attendance requirements. By way of example, in a study of instructional time in several countries, Helen Abadzi of the World Bank, found that school closure and teacher absenteeism eliminated on average 45% of the official school days in a sample of schools in Ghana.

of such a standard, suppose the standard specifies that the attendance record must be completed by 9 am (or before noon, or by the end of second period, etc.) and an auditor arrives at 10 am. If the attendance record has not yet been completed, the school is clearly in violation of the standard and must be disqualified (which in turn results in the disqualification of the number of total schools represented by the disqualified school in the audit sample). Without a specified completion time and without a specification that records must always be kept at school, the teacher can offer any number of reasonable excuses ("I do the attendance at the end of the day," "my bus was late today," "the teacher who usually does attendance didn't come today," "I was working on my books and I left the records at home," etc.) for not keeping attendance, leaving the auditor with no means to certify that the register is not being manipulated and thus calling into question any record-based statistics. As noted, similar standards would have to apply to deal with missing registers ("I forgot it at home today," "another teacher took it home") or missing pages or gaps in the record. And so far we are only discussing attendance, not requirements related to subject mastery, which, if part of a country's completion requirements, would also be necessary to audit in order to prevent a cheapening of the requirements. For the audit to serve its fundamental certifying purpose, it must audit strictly to the agreed definition of completion, without exception, or else the reported figures will not be credible.

Even if an audit verifies that record-keeping standards are being met, it will be necessary to define how the audit will deal with discrepancies in records. To continue the example above, even if the attendance register is complete at the time of the audit, there will often be attendance-related discrepancies that are difficult, time consuming, very costly, or even impossible to address. If the number of students present is less than shown in the attendance record for the same day, the implications for the audit are not obvious. The discrepancy could be due to deliberate cheating (and thus worthy of disqualification), but there are also many very plausible reasons that could apply and these are very common in developing countries. Some children may have become ill and gone home and some may have just wandered off during recess (a common occurrence in schools without boundary walls, or schools without toilets) and not returned. In such an event, does the audit accept the official record (in which case the attendance requirement is reduced to a short appearance at school), is there an allowed maximum level of discrepancy (beyond which the school is disqualified), or does a varying sanction apply?

Education Ministries face a classic principal agent problem in their efforts to control what happens at the school level, and this extends to the maintenance of school records as well. In many developing countries with scattered rural schools Education Ministries have too few supervisory staff and insufficient logistic resources (vehicles and operating budgets) to allow school visits more than once or twice a year. Many of these schools themselves suffer from extreme resource deficits (for example, too few teachers which results in often unmanageable pupil teacher ratios that in places exceed 100) and a lack of incentives that might encourage greater attention to record keeping. In fact, in many places just getting teachers to show up for

work presents a challenge, much less ensuring that they maintain records to some level of standard. These issues of capacity and the lack of incentives and accountability often extend not just to schools but to scattered education offices as well, and thus it may be difficult, if not impossible, for some countries to meet the terms of a contract based on completion simply because they do not and often cannot control what happens within the various levels of the system.

An education records audit would not be limited to school records, but would need to examine the process (both the official and practiced processes) by which school level information is communicated and aggregated at various levels within the country. In some countries regional variations in definitions and record-keeping practices make the aggregation of statistics a very problematic and contentious process. In short, an audit of records and record-keeping practices for the purpose of verifying primary completion numbers is a demanding and complex task.

### 4.1.3. Audit Based on Parent Reports

One alternative that has been proposed to an audit of school record-keeping is to seek confirmation of completion claims from parents. This could be done either through a randomized survey of a representative sample of households or by randomly selecting students from school records and visiting their households. Both of these methods will yield an estimated count, but neither can verify that the definition of completion has not been cheapened or gamed. Indeed, definitional issues (completion as defined by the household vs defined by schools) will likely lead to discrepancies in both of these approaches that are very difficult to resolve. The tendency of some parents to deliberately have their child repeat a grade suggests parental disagreement with automatic promotion policies, and some parents may not accept what they perceive as a cheapened definition of completion. In countries with little or no formal completion requirements the lack of agreement between school and parents is literally impossible to check. In countries that have some formal requirements and issue a primary completion certificate, with the certificate being a formal notification of the child's completion to the parents, this problem could be avoided by requesting to see the certificate. However, if a parent of a claimed completer does not recall receiving a certificate for their child, or is unable to find it, there is no straightforward or reasonably inexpensive way to resolve the discrepancy. (Note that cases where the school claims completion but the parents are not aware are likely to be more common than the opposite, tending to create a situation where school-based reports seem like over-estimates and thus cast some doubt on the government numbers.) We believe the number of discrepancies could be very significant and very expensive, if not impossible, to resolve. Even if the school records did support the child's completer status, without the records audit mentioned above (which the household survey would be intended to obviate) there would be no means to rule out manipulated records. At most these two methods could give an estimate of primary participation that includes a portion of the final year.

### 4.1.4. Attendance Audit Discussion

The examples above illustrate significant difficulties with basing Cash on Delivery contracts on primary school completion. A regular random school audit of school record-keeping over the course of the academic year can capture adherence to definition as well as ghost gaming. However, if a country cannot create a precise definition of completion (and the definition process can be as much a difficult political exercise as a technical one), or if it cannot devise and implement the necessary systems and processes to verify completion, then there are no grounds for a verifiable audit. It is very likely that few countries currently have systems (either in terms of design or implementation or both) sufficiently robust to withstand the scrutiny of an audit that meets the needs of a Cash on Delivery contract. It is also likely that the resources and sustained attention required (both financial and human) to meet audit requirements would distract attention from, and might come at the expense of, efforts to increase access or quality. In environments where capacity is already stretched the benefit in better statistics on this issue may not offset the cost of gathering them. Ironically, it is possible that the audit, and developing the capacity to withstand it, could become the primary focus of the Cash on Delivery experiment and diminish attention to the broader outcomes it is intended to promote.

With regard to surveying parents to verify primary completion, the sample sizes that would be required for a randomized household survey to yield useful confidence intervals are very large since households with children in question are a small subset of all households, or specialized sampling frames will somehow have to be created. The reason for this is that very few households will have children who are likely to have completed that year, and therefore simply visiting households using a standard sampling frame will not work, or will result in a great deal of unnecessary expense. On the other hand, constructing an alternative sampling frame will also be difficult. Combined with the likely discrepancies related to definitional issues outlined above, this method could be extremely costly and of marginal value. More detailed discussion of the issues and possible costs of this option can be found in Annex 1.

If school record-keeping audit approaches to verification of completion numbers are not feasible, and if the donor is willing to risk a cheapening of the definition of completion, a home survey of completers based on school records (not a household sampling frame) combined with an examination of the means by which reported school numbers are aggregated could serve to verify the total numbers, assuming discrepancies can be resolved or that margins of agreement are sufficiently flexible. This type of audit need take no more than a month and could be conducted shortly after the academic year has come to an end. Any firm with national-level survey capabilities could be engaged to perform this type of audit with a moderate level of training and oversight. But, in the end, one would need to question how seriously the Cash on Delivery program would be taken if the audit requirements are so slight.

There is a variety of gaming opportunities (beyond misrepresentation of the numbers) that the completer number audit cannot easily control for. Some examples include lowering the achievement barriers to progress between grades which tends to cause dropping out (assuming the audit controls for attendance), opening large numbers of poor quality schools (with too few and poorly trained teachers, with little support or supervision, with few text books etc.) and dramatically increasing class size in existing schools. It would be very difficult for a contract to include controls for this cheapening of the completion definition without violating the "hands off" principle at the heart of the Cash on Delivery concept.

With either a school-records or home-survey-based audit of completer reports, the contract would need to specify penalties for discrepancies that exceed some agreed threshold. Rather than totally disqualify a country based on crossing an arbitrary discrepancy threshold, the contract could specify payment per incremental completer on the basis of the audit estimate with an escalating penalty tied to the size of the discrepancy. For instance, the contract could specify $100 for each incremental completer, not to exceed the audit-based estimate of incremental completers, and reduced by $5 for each percentage point by which the upper limit of the audit-estimate confidence interval is below the discrepancy threshold. For example, suppose the discrepancy threshold is 5% and the reported incremental completers is 100. A penalty will apply if the audit is confident that the true number is less than 100 minus 5%, or 95. If the audit-based estimate is 85 with an uncertainty of 3, the penalty is applied to the difference between 95 and 88 (the highest number the audit estimates the true number is likely to be), or 7 which also happens to be 7% of the original report. In this case the country would be rewarded for 85 completers at the rate of $100 minus $35 ($5 for each of 7 percentage points below the discrepancy threshold) each. By escalating the penalty as the discrepancy increases, the country has a disincentive to inflate the numbers and hope the audit is on the high side.

### 4.1.5. Notes of Caution

On the basis of the discussion above, one has to question the value of primary completion as an outcome indicator when the stakes are this high. Certifying completion numbers and adherence to definition is not a simple task, and, unless the definition includes an achievement requirement, it at most certifies a child's physical presence at a school over a period of time and nothing more. The assumptions regarding the value of primary participation alone are increasingly being challenged, particularly as scores on skill assessments such as the EdData II Early Grade Reading Assessment (EGRA) among student samples in some countries hint at very low learning standards. The Cash on Delivery concept is at the forefront of development thinking, and in that spirit we suggest that its proponents downplay the emphasis on primary completion and encourage the shift of donor focus towards more meaningful outcome indicators that are tied to quality of learning. In some sense, innovativeness in what the development-aid approach is *applied to* should match the innovativeness of the approach itself. In that same sense, it seems a

waste to focus such an innovative approach on a goal that is at worst of questionable value and at best has extremely difficult audit requirements.

As an added (and we believe important) note of caution, we point out that few (none that we know of) countries have had their existing primary completion statistics subjected to any significant level of formal and in-depth scrutiny by any outside agency. Completion *rates* as understood by international reporting, at this point, is more a matter of estimates, in any case, derived from last-grade enrollment, minus last-grade repetition, divided by population of appropriate age. Thus, the very process of establishing a baseline against which progress in completion can be measured could provoke an embarrassing crisis if it emerges that published statistics (both current and past) differ significantly from reality or that they are only rough estimates based on information with embarrassing discrepancies or gaps. This is not to say that one is likely to discover outright cheating. Current household survey data do match completion numbers. But an audit could uncover just how flimsy the very notion of completion is, in varying ways. It could uncover that many countries have few if any formalized requirements, for example, or that they have some formalized requirements but no certification, or have both but really do little to verify anything, at present. This is a very real risk and should be carefully considered before proceeding down this path. Of course, from a certain point of view, uncovering these problems would be a real contribution, but one has to wonder whether it would deviate from the basic aims of the Cash on Delivery program.

Finally, one possible way out of all these problems would be to propose that while the formal "accounting" audit requirements are so daunting as to perhaps be impossible to meet in any conception of a "hands off" approach, or for a reasonable cost, nevertheless the popularization of these issues, brought about by a Cash on Delivery program, and the requirement that data be published, will unleash sufficient social audit and civil society watchfulness as to control for the complex and subtle problems (e.g., monitoring attendance over the year, ensuring that non-attenders do not complete) that a formal audit cannot detect. In a way this would be more consistent with a true hands-off approach: one simply admits that an "accounting" audit sufficiently accurate to detect cheating is not "hands off" and hopes that social audit pressure will do the job and is more "hands off." One then goes with a very simple audit that basically checks the numbers in a very superficial way, and hopes that social audit pressure will take care of the rest. Whether donor sponsors will support this approach, or whether this sort of admission would undermine the credibility of the program, is something one may need to discuss.

### 4.1.6. Technical "Sampling" Issues, Level of Effort and Cost – Completer Audits

With the various conceptual issues out of the way, we can provide some notion of the technical "sampling" issues involved in auditing completers. We use "sampling" in quotation marks because the issues do not refer only to classical issues of statistical sample size and method.

In all that follows we assume that the unit of observation is the school, not the individual child. One is trying to estimate school-level parameters, not child-level parameters. This is because the issue at hand is not whether given children completed or not, as a binary variable, but the number of reported completers. The school may have invented children, after all, or someone at a level higher than the school may have changed the numbers by inventing children, not just changing the status of a child.

We start with the least-demanding case. Since many of the statistical and practical assumptions needed in this case will hold for the more demanding cases, this first case turns out to be the longest, even though it is the least demanding. The number of paragraphs spent on this case therefore does not imply an "endorsement" of this case, as it is indeed the least demanding and in some sense the least satisfactory case. In this first case the donor accepts whatever definition or non-definition of "completer" the country proposes, or already uses, including something as simple as last-grade enrolment minus last-grade repetition (the current default used by, say, the World Bank), and simply checks the numbers.[4] In fact, if there is no definition or formal certification of completion, there would be no choice other than this default case. If this is judged insufficient, then perhaps the country should not be eligible for Cash on Delivery. (There are technical issues here as basic as some countries not collecting data on end-of-year or final enrolment, and simply using initial enrollment, and subtracting repeaters, to estimate completers. We assume the donor is satisfied by a definition as simple as initial enrolment minus repetition.) All this means that one simply checks the numbers even if there is no formal definition at all, just to make sure there is consistency between what the schools report to the country, and what the country reports to the donors. More importantly no attempt is made to measure whether there is any gaming of the *notion* of completion itself (including sudden easing of repetition criteria to pump up the derived completion numbers as a one-off trick), and the audit simply checks that the *numbers* are not somehow misrepresented along the reporting chain. In other words, anything that one finds at school level is taken as "truth." We will also assume that the donor and the government have no interest whatsoever, if there *is* measured misrepresentation of the data, in determining at *which* level (above the school, namely circuit, district, region, province, etc.) the misrepresentation or manipulation of the data is taking place, as this would be too much of a

[4] See definition of the Primary Completion Rate used by the World Bank's EdStats at http://web.worldbank.org/WBSITE/EXTERNAL/TOPICS/EXTEDUCATION/EXTDATASTATISTICS/EXTEDSTATS/0,,contentMDK:21272698~menuPK:4323930~pagePK:64168445~piPK:64168309~theSitePK:3232764,00.html#o which is, basically, enrolment (including most likely initial rather than final enrolment) minus repetion. The World Bank also reports on the Gross Primary Graduation Ratio, which comes closer to a notion of certified completion. But a much smaller number of countries (13% out of all possible) report on the graduation ratio than report the data needed to calculate the completion ratio (48%). UNESCO's name for the same variable is perhaps more transparent than the World Bank's, in that instead of using a name such as a Completion Rate, UNESCO calls the same concept the Gross Intake Ratio in the Last Grade of Primary. The latter name is more transparent as it lays out the logic of the concept. (The data are exactly the same.)

departure from a "hands off" approach (and would increase the cost of audit)[5]. In this case it seems to us that the simplest audit criterion is, ultimately, to check a random sample of schools.

To drive the sample size calculations for this first case, the following assumptions are made:

1. The parameter to be estimated is the number of extra completers in the country or, to operationalize the concept, the average extra completers per school, μ.

2. The approach would be to create a one-sided 95% confidence interval for the average school-level increase in completers. (This can then be "blown up" using the number of schools to get the relevant parameters for the total.) The interval can be one-sided since one is interested only in ensuring that the country is not below its claimed level. The idea then is that the country "passes" if its claimed per school increase in completers is less than the upper end of the one-sided confidence interval $\bar{x} + t_{n-1,.05} \; s/\sqrt{n}$ where $\bar{x}$ is the sample mean, $t_{n-1,.05}$ is the *t* distribution ordinate for n-1 degrees of freedom that leaves 5% of the probability to its right (about 1.66 for large *n*), *s* is the sample standard deviation, and *n* is the sample size. If the country's claimed increase in completers is above the upper end, it could be disqualified entirely for attempting to falsify (this seems unduly harsh), or it could be given credit only up to the upper end of the confidence interval possibly with a penalty for the overestimate.

3. To fix ideas, we will assume that the standard deviation for increases in completion at schools is 20. This will have to be refined based on analysis of real data for a few countries, we have derived it from reasonable assumptions based on our knowledge and data from one country.

4. If one assumes, further, that the mean increase will be about 5 completers per school (a reasonable value if one assumes that the completing number per school are 50, equal to one session of 50 students or 2 sessions of 25 students). If that is the case, then a one-sided confidence interval width should be 1. That is, the product to the right of the $\bar{x}$ in the expression for the confidence interval should be 1.

5. Based on those assumptions, a sample size of approximately 1,100 schools would be required[6]. Assuming that two schools can be visited per day, this implies some 550 person-days of field work. If one assumes that field work costs $150 per day (including overhead costs), and if one assumes that field labor cost is 1/3 of total cost (total cost including supervision, transportation, per diems, simple analysis, and report

[5] Determining the level where the manipulation is taking place is not necessary for the audit, and is interventionist. However if this determination is useful to the country, the donor could finance it as part of the audit process.

[6] This figure is largely independent of the total number of schools in the country.

preparation), then the cost of the audit would be $240,000. If one assumes that three schools can be visited per day, then the cost would come down to about $170,000.

6. Note that the reason for such a large sample size is the narrowness of the confidence interval, and the perhaps pessimistic assumption on the standard deviation of the yearly increase in completers (large, relative to the *mean* yearly increase). The narrowness of the interval is driven by the fact that one is trying to detect change, not estimate a parameter at one point in time. A one-sided width of 1, when the average change is 5, is actually fairly broad. Before making any firm conclusions with regard to sample size, it would be necessary to investigate what might be the standard deviation of yearly increases. We have opted to be relatively cautions (maybe even pessimistic) at first.

Note that it is much easier, and probably a better audit practice, to measure the average increase across schools, rather than whether the records diverge, school by school, between those at the highest level at which school-by-school data are kept, and those at the school level, where the "punishment" would be proportional to the number of divergences or the average degree of divergence one found, or would totally disqualify the country if more than a certain level of divergence was found. The reason for that is that a zero divergence would "pass" from an audit point of view, regardless of the measured increase. Thus, if a school reported no increase at school level, and the record was the same as reported by the country, the school would "pass" the audit. Another important advantage of simply checking whether the school-level increases match the nationally-reported increases is that one does not need to match records between the school level and the national level (or highest level at which school-by-school data are still kept). If one takes the "divergence checking" approach, instead, and the system knows that the sampling proportion is, say, 1 in 15 (e.g., 1,000 schools in a system with 15,000), the system could react by manipulating a few schools a lot, rather than every school a little, knowing that a sample would be unlikely to catch the few schools that have been heavily manipulated, and that the audit is not checking for a discrepancy between the totals, but just checking to see how many discrepancies there are. Thus, if for some reason one wanted to verify correspondence of records, rather than just overall correspondence of reported levels, then it may be wise to sample schools proportional to reported increase. In that case, it may be important to construct a sampling frame at some level of the system where data in the EMIS are still available by school. Thus, another big advantage of just checking the levels is that no special sampling is required.

One also needs to note that schools would have to be instructed to keep their enrolment and repetition records for at least the past year and the current year, in addition to the requirement that records always be at school by a certain time of day (as noted above), so that increases can be verified. Even with that requirement, there will be some schools that, on the day of the visit, will not be able to produce evidence. In these cases, for those schools, the increase would have to be counted as zero.

A slightly more demanding case would be to suppose that one does not accept "truth" as being that which one simply finds at school, taking school records at face value (i.e, enrolment and repetition as recorded at school level), since schools may be manipulating the data in some crude way (e.g., inventing children, or declaring them to have completed even if in some sense they did not), under pressure from the system. Thus, if one wanted to go one simple step beyond the most basic case outlined above, one could ask to talk to the family of a random sample of the children at each school. This would be feasible, but it would multiply the cost by, perhaps, five to ten times, resulting in a cost closer to $1 to $2 million if one kept the same number of schools. The time required to visit, say, five to ten families per school would be considerable, meaning that each school might take several person-days of work, instead of one-half person-day. Furthermore, the sampling would become considerably more complex, as it would now be a cluster sample. On top of all this, it would not be clear what one gains, because if the families disagree with the schools, it is difficult to prove who is right without considerable further work, and there is no *a priori* reason to believe the family more than the school, particularly in countries where the criteria for completion, and the definition of completion, are very loose. Even if there was a certification process, the family may not remember that the child received a certificate. The one thing this process could catch, unambiguously, is completely invented children.

One twist on the school-plus-home-visit option discussed immediately above is to undertake a random sample of households. As already noted, it is unclear what this would achieve, and the cost would be high. Few households will happen to have a recent completer, so one would have to create a special sampling frame, presumably—a very costly undertaking. And, again, if there is no certification based on a clear definition of completion, it is unclear how one resolves differences between the household-derived data and the claims the country makes based on school records, particularly at an aggregate level, since in this case there is no possibility of going back and forth between school and household.

The most demanding, and perhaps most useful approach can be used if there already is a strict definition of completion, hinging on attendance, for example, or other considerations. In this case the sampling audit approach would check to see both whether the system is falsifying school records, and whether school records needed to justify some form of "certification" of completion or graduation are: a) properly kept, and b) do in fact justify. The simple fact-checking aspect is already included in the "least-demanding" scenario already discussed, so it is not covered here. The more demanding aspect is to assess whether the school's records are being properly kept and justify the issuance of the completion or graduation certificate, or simply recording and reporting the child as a completer or graduate (if no actual certificates are issued). To do this, we would recommend that a (different) random sample of schools be visited at random several times during the year. The behaviors to be checked would include all those that pertain to pupil flow between grades, attendance issues, and so on. The checklist would assess

both whether students proceed based on the various criteria, and the quality of the school's record-keeping. A checklist would either qualify or disqualify the school. The checklist would have some allowance for errors and omissions, but errors and omissions in record-keeping beyond a certain level would disqualify the school as counting towards the completion numbers, resulting in a penalty.

As already noted, one issue that needs to be considered in this approach is whether one might discover that a country really has not been keeping its records, and therefore the past completers or graduations are all suspect. One way out of this dilemma is to consider the quality and veracity of record-keeping discovered in the first audit as a baseline, and then disqualify any increase in shoddiness (motivated by the pressure to increase completion) from the baseline.

The yearly cost of such an audit would probably be five to ten times larger than the cost estimated for the simplest scenario discussed above, or $1 to $2 million.

These costs have to be put into perspective. Current donor projects that affect only a small portion of schools, and do little to systematically improve record-keeping or accountability (the information systems components of projects are not often sustained) frequently cost $20 to $100 million, in just one project. To look at it from an overall systems perspective, the average annual primary education expenditure in low and middle income countries is approximately $5 *billion* (in low income countries it would be less). Thus, this audit would typically cost less than 0.021% (that is, about 2 percent of 1 percent) of the primary education spending. This seems like a low price to pay. On the other hand, if a country's participation in this scheme is yielding only some $20 million in *additional* resources for the country, spending $1 million on verification does seem excessive, at 5% of the cost. Thus, the verification has to be explained (if one is interested) as a way to improve functioning of the overall system.

## 4.2. Test-takers

### 4.2.1. Advantages of a Test-Taker Measure

Basing a Cash on Delivery contract on the numbers of students that take a standardized test adds a quality dimension to the incentive that is missing from the attendance-based completion indicator. Testing of children at the end of primary school is a good thing in and of itself for many reasons. It allows a country to assess the return on its investment in education (at least in terms of learning) and introduces an element of accountability at all levels in the system. A Cash on Delivery agreement could reward annual increments in the number of test-takers in addition to payments for increments in attendance-based completion, or test taking could itself become a measure of completion as testing becomes more widespread. Or test taking could be the indicator of choice, ignoring completer numbers.

An audit to verify the number of test-takers is, we believe, potentially much easier to carry out (leaving aside the topic of test scores for the moment) but it is still not as simple as it might initially sound. It must, for one thing, control for who is taking the test. It must be able to identify double counting or the inclusion of fictitious test-takers and must also control for real but ineligible test-takers (adults, out of school children, primary school graduates from prior years, children from neighboring regions etc.). However, even with these caveats, the number of test-takers is likely much more auditable, because testing is an event fixed in time, with a totally clear definition, without requirements that are spread out over time. One is simply auditing appropriate presence at a single yearly event, as opposed to testing whether year-long (and in effect, career-long) requirements for completion were met.

### 4.2.2. Test-Taker Audit Requirements

To achieve control over numbers, it would be necessary for each test-taker to have a unique identifier. A national identification card number, common in many countries, would be perfect as it does not impose the burden of assigning numbers on the Ministry of Education. Such a system assigns an ID number to every person, not just each student, and it has controls against assigning more than one ID number to any one person. The ID number would have to be matched electronically in a database containing test-takers over time, in order to prevent the same test-taker being counted twice in the same year or counted again in subsequent years. This ID number would eliminate the incentive to include children from lower grades as they would not be able to take the test in subsequent years (when they are properly eligible and expected to perform better). It would also prevent students who have taken the test at the end of primary school from taking it again in subsequent years. Ideally the ID numbers of all primary school completers (not just test-takers) should be included in the database to prevent the inclusion of untested primary graduates from previous years in test-taker numbers. If this is not done, until test taking becomes very widespread, the inclusion of untested primary school completers from prior years represents a short-term gaming opportunity that is not easily controlled for. (After some time it would not be possible to game the test numbers in this manner in any case.)

We propose that an audit of test takers be based on a retest in a random sample of schools (truly random schools, including schools that are difficult to reach, in order to prevent gaming in schools considered unlikely to be visited or in fact known to be out of the sample—that is, the sample would not be able to rule out schools that are "too far" as is often done in random school samples drawn only for research purposes) very shortly (one week, perhaps) after the official test. In order to maximize attendance on the day of the retest (and therefore prevent excuses regarding selective entry in the re-test, which could mask cheating on the test itself), the date of the retest (though not the identity of schools in the random retest sample) would be pre-announced. In addition to controlling for cheating (preventing cheating to increase test scores is not a necessary audit function for certifying exam takers, but an audit that reports cheating has an

important social audit function) the retest audit has a number of advantages over an audit on the day of the official test. Apart from administering a retest, the audit team would compare retest numbers with the test-taker count and would be required to verify the identity of test-takers who did not appear in the retest (which would identify adult test-takers and facilitate identifying out-of-school or out-of-region children). The audit team would also compare test-takers with the enrolment list. A comparison of scores (which could include item correlation) from the test and retest would help eliminate out-of-school children (since their scores would presumably be very low on the retest). The retest also greatly increases the cost of gaming by importing test-takers from neighboring areas since imported students would have to be physically present on the day of the retest in all schools.

.

One gaming strategy that the retest exercise cannot adequately control for in the short term is the inclusion of young primary school completers from previous years whose ID number is not in the database (although to fool the audit, the school would need to include these test-takers in their enrolment records). This is only a temporary problem, though, since immediately following the introduction of the ID number tracking of primary school completers, each primary completing cohort is prevented from appearing in the test in subsequent years, and schools are likely to be reluctant to "use" a child only once *and* early, as this would bring down their scores without resulting in a permanent increase in the number of test-takers. As the age differential between those whose primary completion is not registered in the database and the expected age of test-takers increases, they will be more easily identified by the retest audit (primarily by visual inspection since birth records are unreliable in many countries and in any case the age of students in primary classes can vary considerably depending on age at enrolment and repetition). The audit would also have to ensure that records from the original test are not manipulated after the retest has taken place (otherwise ineligible tests could be added to the count from schools outside the retest audit samples).

In the absence of a baseline for test-taker numbers, the Cash of Delivery agreement could specify a fixed sum payment for the first year and use end-of-first-year figures as the baseline for future payments.

Apart from the cost of the audit (which we investigate below), use of the test-taker indicator would require significant investment in a range of necessary systems. These include the design and implementation of a system to allocate ID numbers (many, though not all, developing countries have national ID card systems already), setting up an electronic student database, test development (which may require establishing a national testing agency) and the logistics necessary for widespread and secure test administration (including systems for scoring and reporting).

### 4.2.3. Test-Taker Audit Discussion

There are some significant advantages to using a test-taker count as an indicator of progress, both with regard to audit feasibility as well as from a development perspective more generally. The parameters of the audit in terms of tasks, scale and duration can be precisely defined and the opportunities for manipulation and cheapening are more easily monitored compared to primary completion measures. At a broader level, the testing capability that this indicator requires gives a country the means to monitor the quality of education (the government and civil society can look at test scores, even if the Cash on Delivery contract does not require it but requires only that they be published) and thereby target educational resources more efficiently as well as establish forms of accountability that have often proved elusive in the education sector. From the donor's perspective, the test-taker indicator can be measured with greater confidence and probably less effort, the indicator can provide a better measure of the desired outcome (more test-takers scoring more) since test scores would be available even if not factored in to the contract, and the indicator has a built-in potential to facilitate improvement incentives (measures of test-takers and their scores can be used as the basis for rewards and sanctions within the system).

In the discussion above, we have deliberately avoided a recommendation on how to treat test scores. Including test scores in some fashion as a condition for Cash on Delivery leads to a wide range of problematic issues and perverse incentives.[7] Any merits related to the relative advantages of trends in averages versus pass rates would be complicated by existing public perceptions regarding the use of high stakes examination scores, and the difficulty arises of what to do as test scores fall due to introducing the tests to more marginalized school populations. Rather than try to navigate through this complex set of issues, we propose that the contract be limited to rewarding total test-takers but require the public dissemination of test and audit results, disaggregated to the school level and maybe according to guidelines governing presentation that facilitate understanding of trends in quality and equity. This would facilitate public discourse on a range of topics related to education quality and equity that would serve a social audit function and encourage the government to respond to issues of concern to civil society rather than those of interest primarily to the donor. This social audit function is one of the core strengths of the Cash on Delivery approach.

Audit reports, in addition to test scores, that would facilitate the social audit function include names of schools whose retest-taker count is below the claimed test count by a threshold percentage (which might suggest ineligible test-takers), schools whose retest averages are below the test averages by more than a threshold percentage (to identify cheating), schools whose test count is below enrollment by a threshold percentage (which might indicate selective test-taking in which weaker students are prevented from taking the test) or schools whose penultimate grade enrolment is below the final year enrolment by a threshold percentage (which might indicate

[7] For a fuller discussion see "Informal reflections on audit issues surrounding "Progress-Based Aid" in the education sector" by Luis Crouch

selective promotion or "push-out" to eliminate weaker students). The donor-funded reward for test-takers should function as a balancing incentive against selective test-taking to improve scores. The retest feature of the audit is a critical element of the social audit function and should not be abandoned for reasons of cost or convenience.

The topic of testing invariably invites debate regarding teaching to the test. Teaching to the test is a spectrum, and assuming the test is a good measure of intended learning outcomes some of it is desirable to the extent that it keeps teachers on task. Taken to extremes it becomes problematic, but the social audit function of publishing test scores should allow this issue to work itself out without intervention by the donor. Not making actual scores a subject of reward, but simply requiring that they be published, should also lower the stakes somewhat, and reduce undue pressure to teach to the test.

If test-taking is to be a recommendation as a progress indicator for Cash on Delivery a few notes on the test are in order. It is not uncommon for national primary curricula in many countries to be rather ambitious given the resources available at the school level (in terms of teacher numbers and qualifications, materials including textbooks, and physical plant). In many cases textbooks do not cover the entire curriculum and frequently teachers are unable to cover the entire prescribed text over the course of the academic year. Or there are not enough textbooks, or they arrive too late in the year. There is thus a high likelihood that few students have been exposed to, much less have mastered, the entire curriculum, and a test developed to measure learning outcomes specified for the final primary grade may well result in very low scores for many or most students which would be discouraging for all concerned. It would be better for initial tests to focus on a core set of key (relatively curriculum-independent) skills in a limited range of subject areas such as reading and math and to include items covering a range of grade-specific levels of difficulty to better discriminate between the variety of achievement levels that may be represented in the final year of primary school. The use of multiple choice tests would reduce problems of consistency in scoring and would allow for item analysis. In countries where testing is not currently practiced, there would need to be significant investment in developing the capacity and infrastructure necessary for wide scale testing. Accepted approaches to test development that ensure psychometric comparability from test to test as well as systems and logistical requirements necessary to ensure test security and smooth implementation can be readily adapted to meet the needs of individual countries.

### 4.3. Technical "Sampling" Issues, Level of Effort and Cost – Test-Taker Audits

The technical requirements for auditing test-taker numbers and scores would be similar to the most involved of the completer audit ideas. It would imply a random visit to a number of schools, re-testing the children, and assessing the total numbers to make sure "ghost children" or "ringers" (capable children from some other school, or who already graduated) are not taking the

test. A similar confidence interval approach to that discussed above could be used for the scores. If the overall claimed or reported scores are higher than the upper bound of one-sided confidence interval derived from the re-test, the testing would be considered fraudulent. The size of the confidence interval should be set at about 5% of the average score, and the sample size needed to calculate this number could be used. To fix ideas, we use data from international tests. We assume an average score is 250, with an SD of 100. To create a one-sided 95% confidence interval for the average school results with a width of 12.5 (5% of 250) would then require a sample size of only about 180 schools (assuming one tests all of the final year students, some 9,000 students—a number that roughly lines up with the numbers used in sample based testing in most countries for tests such as PISA). Note that, therefore, the confidence interval at the student level would be much narrower than the confidence interval at the school level—we assume that the school is the unit of analysis of interest. Now, 180 is much lower than the 1,100 we somewhat pessimistically suggested for the audit of completer increases. This is because in the case of the testing one is trying to establish only whether the re-test score is within the confidence interval of the original score, and not whether *increases* are within the range of each other. However, the cost per school would be much higher than is the case with the simplest audit discussed above, because it would most likely take 2 person days per school of field work, not ½ of a person day. Furthermore, there is probably more analysis involved. All in all, the numbers might be ¼ as few as in the simplest audit above, but the cost per sampled unit would be about 4 times higher, leaving the total cost about the same: somewhere around $250,000, perhaps. (This figure probably deserves more thought. The cost might be low relative to other testing exercises, but note that presumably the testing system itself would have worked out all the fixed costs of the technical testing details and testing infrastructure, so that the costs of the audit would be largely marginal in nature.)

### 4.4. Secondary Enrollment

Another indicator under consideration as a progress indicator in a Cash on Delivery contract is secondary enrolment. This incentive has a variety of serious weaknesses. Secondary access (both physical and economic) is already a problem in most developing countries, and increases would almost certainly be at the expense of quality. Also, in many countries there is already a surplus of primary completers who do not have access to secondary school, so an increase in secondary enrolment would not require an increase in primary completion. Furthermore, such an indicator introduces very perverse incentives such as providing access only to the first year of secondary school—that is, enabling a lot of entrance and then quickly pushing out the youth, creating a skewed grade distribution and more dropping out. Finally, secondary school enrolment is already growing faster than primary school enrolment even in low income countries, so it is not clear that further incentives are needed. Beyond all this, the audit required for such an indicator is even more complex and problematic than what is required to verify primary completion.

## 5. Recommendations

Based on the options currently under consideration we offer the following recommendations:

1. If rewarding increments in the number of primary school completers is absolutely necessary and the donor wishes to avoid the more complex records-based audit, namely the donor is willing to accept as "truth" whatever is happening at school, the donor could opt to conduct a random sample survey of school records. This would show very little, though, since it would have to accept as truth whatever is found in school records. Bringing in data from the home, via a household survey (either linked to the school or not) could add some additional information, but how one resolves differences between home-based data and school-based or system-reported data is not at all clear. Neither of these options would control for a cheapening of the definition of completion and thus risks discrediting the entire Cash on Delivery approach.
2. If rewarding increments in the number of primary school completers is necessary, and the donor needs confirmation that the agreed definition of completion is being consistently applied, the donor could take on the complexities of the records-based audit, acknowledging that it is much less hands-off and could have embarrassing implications by revealing how poor the baseline estimates really are (and we believe they are generally rough estimates at best). However, the whole process would be most beneficial, and there may be ways of dealing with the embarrassment created if one discovers that schools are not following procedure, by focusing on improvements from the baseline.
3. The donor could change the progress indicator to a count of the number of test takers (combined with mandatory reporting of test scores and audit findings) and use this either as a stand-alone indicator or as a proxy for completion. In this case the donor may wish to increase the reward for test taking since $20 is not enough of an incentive to encourage an increase in completion *and* test-taking. Performance on the test would not be subject to reward, but performance would have to be published and audited, and tampering with the performance data could lead to partial or total disqualification.

The 10 year agreement envisaged under the Cash on Delivery proposal is a sufficient time frame to consider several implementation options based on Recommendation 3 above. The agreement could set a 10 year goal of universal test-taking and reward increments in test-taking at higher and higher levels as coverage approaches 100% (since increments at that stage would be due largely to incremental completers). Alternatively, the contract could reward all test-taking at an amount lower than $20 to cover the unit costs of testing (at least in the early years) and reward annual increments over the previous year at a higher level, say $40 or $50 and raise this to $100 when testing coverage exceeds a target (such as 75% of all completers) since incremental test-takers at that stage would start shifting to rewarding incremental completers. In this way the

test-taker indicator would serve as a tool to achieve universal completion *and* testing, something that would be a significant achievement indeed.

## *Annex 1: Some Notes on a Household Survey*

In order to estimate completers based on a randomized household survey, it is possible to calculate some hypothetical figures on what sort of household sample size one might need, and what the cost would be. But in order to anchor these calculations better, and make them more meaningful, one needs to think about the type of household sample that might be suitable, and why.

First, completely aside from any considerations already discussed, it is unlikely that one could use a household survey to estimate absolute number of actual completers. A sample could indeed be used to generate a sample-based count of completers, but in order to expand this to a national estimate, one needs expansion factors or weights. But to get these, one has to have a fairly accurate estimate of the total population, or universe. The key issue here is not that one needs sample weights so as to properly calibrate the influence on, say, averages or proportions, of the different types of families according to, say, geographical location. The issue is not proportional weights to take into account different degrees of representativity of different households chosen in a clustered or stratified fashion. The main problem is using the weights to aggregate back up to a total population, when the total population itself is a matter of considerable uncertainty. If a very good population count and sampling frame is not available, then samples are still reasonably valid for estimating means and proportions, but their validity for calculating absolute estimates of totals, with the levels of precision required by a Cash on Delivery approach, is quite compromised. In any case, population totals vary year on year and are estimated, for inter-censal years, using population projections, not samples. Countries do not rely on samples to generate inter-censal estimates of population precisely because the fact that population is not known means that the expansion factors are unreliable. (Samples might be used to calibrate the population projection, e.g., to derive information on mortality and fertility trends, but samples are not used to actually calculate the population, demographic projections are typically used instead.) In fact, in some countries demographic projections are used to re-calibrate the sampling weights needed to expand sample results back up to the population. For example, if using the expansion factors or weights on a sample produces a population estimate lower than the demographically-projected population, then the sampling weights are adjusted so that the sample blows up to the projected population.) Thus, using household samples to calculate total, absolute, numbers of completers is almost certainly not a good idea in low-income countries.

If one wants to have a household-based estimate of completers, it is probably a wiser idea to first agree on a projection of youth of completion age, for example, youth who are 13 years old. This projection would have to be agreed to exogenously, or as part of negotiations, prior to making

any commitments. This projection should be "smooth" with regard to the population base, for the same age group, that would be used in the baseline to estimate a completion rate or ratio; that is, there should not be any discontinuities in the population projection around the baseline estimate.

To better fix ideas, one could propose to measure the completers, based on household surveys and a population projection, as

$$\hat{C}_t = \hat{c}_t \; P_{t,13},$$

where $P_{t,13}$ is an exogenous and previously agreed population projection, and $\hat{c}_t$ is a proportion estimated from a household survey. One would know that there is a $C'_{t=baseline}$ estimated not from a survey but from administrative records, and hence with an unknown confidence interval. For $\hat{C}_{t=baselineyear}$ one would have to ensure that $C'_{t=baseline}$ is inside the confidence interval for $\hat{C}_{t=baselineyear}$ and similarly for the years beyond the baseline. The reason for requiring that the numbers agree—within width of the confidence interval—is obvious. But it is also key that the administrative records and the records estimated from the household surveys agree in the baseline, otherwise the projections will necessarily also be wrong, in a way that could have been predicted ex ante, and this could become a contentious issue.

However, a cost issue arises with this procedure. The proportion of completers *c* is relative to a denominator consisting of 13 year-olds (or whatever the appropriate age is—this would have to be fixed in advance). But in most countries the number of households with 13 year olds is a very small proportion of the total number of households, perhaps around 10%. This means that to get a meaningful confidence interval precisely for this population group, if one relies on a standard household survey where households are not selected based on their having 13 year olds, one would have to have a large survey indeed. For example, assuming that about 10% of households would have 13 year-old children (a reasonable assumption on the low side), and assuming a completion rate of about 50%, a household survey of 30,000 would be able to establish a confidence interval on the graduation rate of [.478, .516]. With a population of 13 year olds of 500,000, which is what one would expect in a country with a total population of 15 million or so, that is a total (two-sided) margin of 19,000.

If the country in question already has a *yearly* household survey of this magnitude, then adding a question regarding whether any child in the household finished primary school the previous year (or the same year, depending on whether the survey takes place after the end of the school year, but within the same calendar year) would be very inexpensive, if it would cost anything at all. If the survey was not a yearly survey, then the donors would have to decide whether they would be happy with the periodicity of the survey.

However, in either case one would need to note that the quality of the survey would have to be audited, since presumably there is still interest in the numbers being audited by a third party, and these surveys are carried out by governments themselves, or contractors who have might not be fully independent.

If there was no existing survey that was large enough, one might have to carry out a special-purpose survey—but this would be very expensive, probably too expensive if the sole purpose was to confirm completion. It might appear that a less expensive alternative (if there were no existing large household survey) would be to use some form of cluster sampling to select clusters, then enumerate all families within the clusters, and create a local sampling frame in each cluster consisting only of families with 13 year-olds. But this would be a deceptive economy if the only question in the survey related to the completion issue. Going to households to establish the presence of 13 year-olds and then carrying out a survey among households who do have 13 year-olds would actually require more labor than simply asking the few needed questions in the households that do have 13-year olds. If one knew ahead of time the proportion of households who have 13 year-olds (using 10% for now as an example), one could simply take, say, 100 clusters, random-sample 300 households in each, visit all 300 quickly, ignore the ones without 13 year-olds, and ask a few questions of the ones who do have 13-year olds, and one would end up with about 3,000 13 year-olds, to get a fairly narrow estimate of the completion rate. A very back-of-the envelope estimate of the cost of such a survey would be around $500,000.

It needs to be emphasized yet again that all this would establish is whether the numbers reported by the households to a third party auditor match the numbers reported by the government from administrative records or possibly from their own surveys. There would be no information on the quality of the completions, whether the completers are even attending or are simply being declared completers, on the source of discrepancies if any, and so on. If there are discrepancies beyond the confidence interval, one would furthermore not know whether they are due to sampling error or other forms of error, such as conceptual disagreements between households

and schools, or, indeed, outright misrepresentation along the reporting chain from school to circuits or districts and so on up to the national level.

As already noted, one can combine a household study with a school study, but using the school as a primary sampling unit, and then visiting some of the households that belong to a school and which the school claims have completers. This would do away with the problem of finding a sampling frame for 13 year-olds, but would imply adding a cost to the cost of a schools audit. The additional cost has already been discussed.