Assessment of Multilingual Education Programs

Stephen L. Walter and Mirinda Burarungrot

1 Introduction

Assessment in education[1] is a massive industry. Worldwide, billions of dollars are spent annually in a vast range of assessment activity ranging from simple classroom assessments designed by the classroom teacher to massive international assessments such as TIMSS or PIRLS. In the US alone, it is estimated that standardized testing costs at least 1.7 billion dollars a year (Chingos, 2012). Compared to such a massive level of investment, a single paper of limited length can cover only some small corner of this domain of assessment even when that domain has been limited to the topic of multilingual education in Asia.

Three further considerations have guided the writing of this paper. First, assessment designs can be as simple as asking an individual a simple question or so complex as to require highly sophisticated designs and cutting-edge statistical methods to analyze data. Second, practitioners or would-be practitioners of assessment work may have zero preparation for such work, or may have PhDs in assessment along with thirty years of high-level experience. Third, many of the readers of this paper will not be native speakers of English.

Accordingly, the authors of this paper have chosen to write the paper for a very specific target audience. We accept that this choice means that other target audiences may find the paper a little difficult while others find it overly simplistic. We justify our choice by reasoning that those who would find the paper overly simplistic already have the knowledge and skill to design appropriate assessments. Those who might find the paper too difficult are not likely to engage in the level of assessment activity described in the paper.

Our target audience consists of mid to upper-level professionals who are involved in or have an interest in multilingual education but lack training in program assessment. This includes both national and international personnel assigned to or responsible to their sponsoring organization or their government for program activity. We are not specifying that such individuals be educators though we assume some level of working knowledge in the field of education. Accordingly, this paper has been written more as a practical guide than as a carefully argued and documented academic paper. Given both its length and its character, we have chosen to limit the number of bibliographic citations and references in favor of practical utility to program managers

In the sections which follow, we describe, in considerable detail, the content and process of program assessment. At selected points, assessment is described at two different levels—basic and more advanced— and this will be indicated when it happens.

[1] The term “assessment” is widely used and applied to a range of evaluative activities with respect to education. In this paper, we apply the term primarily to investigations of the comparative educational effectiveness of experimental programs in mother tongue-based multilingual education relative to national language models. Since student learning outcomes serve as a common metric for such investigations, the methodology described in the paper focus on this approach to program assessment.

2. The fundamental logic of assessment

The basic logic (or process) of assessment is quite simple in a summary characterization: (1) state a question which needs to be answered, (2) identify the information or data needed to answer that question, (3) develop a strategy for gathering the needed data; (4) analyze the data in such a way that the initial question is answered, and (5) present the findings to all relevant parties. While the process of assessment seems simple, experience has shown that none of the five steps, not even the first one, can be assumed to be easy to carry out. The process of assessment described above is essentially the same as that of formal or scientific research. In that sense, assessment can be considered research and we will use the term research frequently in the descriptions and guidance which follow.

Most of the remainder of the paper is structured according to the fundamental logic of assessment. In each section, we illustrate the assessment step being discussed using a project carried out in Thailand by one of the authors (Burarungrot). The context is that of a large Malay-speaking population living in the southernmost provinces of the country. Upon entry into school, most children do not speak Thai, the national language, and the official language of instruction throughout the country. For a number of years, the linguistics department of Mahidol University in Bangkok has been leading an experimental multilingual education program in selected schools in the area. In this program, local children begin their instruction in Malay including the acquisition of literacy in this language. Oral instruction in Thai as a second language begins in Grade 1 in preparation for an eventual transition to Thai as the primary language of instruction in later grades. A local university is participating in the project by developing a special teacher training program for local teachers and carrying out some of the assessment activity.

3. First step: Stating the assessment question(s) which need(s) to be answered

The first conceptual step in planning an assessment is to state the research question.[2] The research question drives or controls the data needed to answer that question. Typically, the statement of the research question entails guidance as to where the data will come from and who or how many individuals should participate in the assessment process.

3.1. Who specifies or states the research question?

The answer to this question is not always obvious. Typically, the one specifying the research question is a stakeholder in the program such as a donor, government entity (either local, provincial or national), or program manager. When there are multiple stakeholders, there may well be multiple statements of the research question or multiple research questions. The person or team planning the assessment must

[2] In actuality, there are a number of practical steps which need to be taken before one can begin even considering the prospect of doing an assessment. Funding needs to be arranged as assessment costs money to implement. Permissions of various sorts need to be arranged for. Agreements or contracts have to be worked out as to timetables, deliverables, responsibilities, etc. This paper will assume that these practical details have been worked out.

There is also the issue of policy on human subjects research (HSR). HSR policies vary from country to country. We can only remind readers and investigators that plans for assessment and reporting of results need to take note of the relevant HSR policies which apply to their work in a given context.

consider the questions posed and negotiate with stakeholders the final formulation of the research question(s).

3.2 Refining the research question

The primary problem to be resolved when trying to define or refine a research question is that of resolving terms which are too vague. Here are some examples:

a. The XYZ program is a good program. (What does 'good' mean? What data or what finding will indicate that XYZ program is 'good?' How do we recognize or measure a 'good program'?)

b. The XYZ program is producing better learners than the government program? (What does 'better' mean? What differences in results are needed to determine that the XYZ program produces ***better*** results than the government program?)

c. Children and their parents are pleased with the XYZ program. (What does the word 'pleased' mean or imply? What evidence can we gather to indicate that children and parents are 'pleased' with XYZ program?

d. Children in the XYZ program are less likely to drop out of school? (How much lower do dropout rates have to be to support this research question? Is transferring to another school the same as dropping out? What about children returning to school after being out for a year; are these dropouts?)

When a research question contains terms which are vague or ambiguous, work needs to be done to negotiate a higher level of precision or specificity in the research question. Question (a) is simply not a good research question and needs to be replaced. Question (b) can be improved by giving a more precise definition of 'better'. Ex. Children in the XYZ program will score at least 20 percent higher on the measure(s) of learning outcomes developed by the assessment. There is still a little vagueness in how the measure of '20 percent higher' is going to be derived. Does it mean all children? Most children? A mean score? What about children who came into the experimental program a year late or two years late? Question (c) has the same problem as question (a) and question (d) has the same issue as question (b).

3.3 Working with multiple research questions?

It is common for multiple research questions to be posed. Assuming that the multiple questions have been established as valid and workable research questions, it is up to the assessment team to deal with the implications of having multiple questions. In the best possible scenario, the multiple questions can be answered by means of the same data set. This is most likely to happen when there is a 'major' or overarching research question with the other questions being subsidiary questions. This is the case in the following sets of questions:

Set A:

The children in the XYZ program with score at least 20 percent higher on the two instruments used in the assessment

Female students in the XYZ program will score at least 20 percent higher on the same assessments

Set B:

The children in the XYZ program will score at least 20 percent higher on the two instruments used in the assessment

Children in rural schools in the XYZ program will demonstrate scores within 5 percent of children in urban schools in the XYZ program.

In other cases, the proposed research questions are not closely related. Consider the following set:

Set C:

The children in the XYZ program with score at least 20 percent higher on the two instruments used in the assessment

Teachers trained in summer programs will be just as effective (achieving the same learning outcomes among their students) as teachers trained at the National Normal School.

There may not be enough teachers of one category or other in the XYZ program to answer this question so that additional data would have to be gathered from somewhere else.

3.4 Another major research question is posed AFTER data gathering has started (or been completed)

Once the research questions have been settled, the next step is to develop a plan for gathering the needed data. Not uncommonly, an assessment team will encounter a situation in which, after developing a data-gathering plan, someone or some entity poses another research question which may be hard or impossible to ignore. If the newly posed question can be answered with the existing data set or the data set currently being gathered, the new question does not pose a serious problem. However, if the data needed to answer the new question is not being gathered, the assessment team faces a serious challenge. A simple, "No, we cannot change our plan," usually is not acceptable. Accordingly, we suggest the following two-step process for handling this situation. Step 1 is to prepare a quick estimate of the time and resources needed to gather the additional data and submit that estimate to the person or entity making the request for consideration and response. Frequently, the person or entity submitting the late research question did so assuming no additional time or expense. Once it becomes evident that the new question entails time and expense, the question is often withdrawn. If, however, the person or entity proposing the new question is really serious, they will invite a more detailed estimate or budget of the additional costs entailed in gathering the required data. That leads to Step 2, the submission of a more detailed proposal including especially a revised budget. If the one making the request can cover the cost and has the standing to give urgency to the request, then the assessment team needs to make plans to expand or extend data-gathering as needed to answer the late question.

3.5 Specifying the assessment question: an example from Thailand

In this case, the assessment work done was carried out as part of a doctoral program at Mahidol University. As such, the assessment or research question could be highly focused, "Does first developing writing skills in Malay facilitate the subsequent development of similar skills in Thai in early basic education?"

The research question was chosen to satisfy a number of stakeholders. Obviously, the researcher needed a question which was manageable for a dissertation project. Thai education officials needed to know that the experimental program was making a positive contribution to the educational development of local children. The local community was concerned that their children not abandon local values and beliefs, yet wanted their children to learn Thai well-enough to have access to

opportunities in the Thai-speaking world. Project donors wanted to know that the experimental project was producing results worthy of their investment.

The next section will reveal how the researcher defined the construct of writing in order to identify the data needed to respond to the research question.

4. Step Two: Specifying the data which need to be collected

The data needed to answer a given research question are closely linked to the research question. For example, if the research question were, "Has reading skill development among Grade 3 students in Bangkok improved in the last 5 years?" one would not set about gathering data on the reading skills of Grade 3 students in Jakarta, nor would one gather data about the reading skills of university students in Bangkok. But what about data on the reading skills of Grade 2 students in Bangkok, or data on the professional development of teachers who teach reading in the primary schools of Bangkok? The assessment team has to consider whether or not data closely related to the central questions actually helps answer that question.

4.1 Levels of data

A 'simple' research question in education (and in most social science domains) is often not nearly as simple as it appears. Those who posed the question typically want to know more than a simple direct answer to the question. They also want to know *WHY there was improvement*, or *Under what conditions was there improvement*? So, while a direct answer to the research question may require only a simple set of data, a full answer to the question requires a broader set of data to address the causal factors entailed in the question.

The challenge for the assessment team is to attempt to identify the 'real questions' being asked so that the data gathered can answer these 'unspoken' questions as well as the direct question(s) which guide the assessment. In this paper, we approach this problem by identifying and describing what we will refer to as levels of data.[3]

In the following section on levels of data, we will make reference to the following possible research question, "Are students participating in the XYZ multilingual education program scoring significantly higher than similar students in comparable government schools?" The discussion of each level of data will be based on this hypothetical research question.

4.1.1 Directly specified. The specification of a research question typically makes it clear what data are needed to answer the question. The sample question stated above quite explicitly requires assessment of children in both the XYZ program AND children in similar government schools. While the question mentions 'similar students' in 'comparable government schools', the interpretation of these phrases is left to the assessment team. Does it mean just in the final year of the experimental XYZ program or should every year be included? Does 'comparable schools' mean comparable in size, comparable in

[3] It would also be reasonable to suggest a strategy of requesting that those posing the research question(s) think through what it is that they really want to know so that the assessment team does not have to guess or, worse yet, overlook something that the requesting entity really wanted to know. The reason we are not giving more weight to this option is that we have found that requesting entities may not be able or willing to take this step. Rather, this is often considered to be one of the tasks of the assessment team.

location, comparable socially and economically, comparable in terms of the age, experience and training of the teachers, or comparable in some other way? Sorting out some of these questions leads directly to second-level data—that which is directly entailed.

4.1.2 Directly entailed. Entailed data is that about which there would be a consensus that such data is needed to properly answer the research question even though none of it is directly specified in the research question as stated. For example, 'similar students' means children of the same age and grade in both sets of schools. It very likely also means children of the same or both genders. Since the underlying focus of the experimental program in question is language or language and culture-based, 'similar' would normally imply a comparison of children speaking the same L1 language—some attending the experimental schools and some attending government schools.

'Comparable government schools' would logically mean those attended by similar students, those located in similar socioeconomic areas, and those having an educational history or track record similar to the experimental schools. Finding comparable schools can sometimes get a little tricky. For example, the presence of an unusually skilled teacher can produce skewed results even if other metrics of similarity are present.

Occasionally, one must deal with subtle or hidden variables which distort results. In an assessment carried out in the Philippines, for example, one grade in one school showed results which were inconsistent with those shown in similar schools. It took several hours of interviews before the assessment team finally discovered that a group of transfers from a satellite school had unexpectedly been "dumped" into the experimental classroom who were totally unprepared for the educational demands of that classroom.

4.1.3 Domain motivated. Data from this level typically refers to data which the assessment team believes important to include based on their professional experience or their knowledge of relevant effects reported in the professional literature. Examples include data on SES[4], teacher skill, availability of textbooks, parental engagement with children in the development of reading skills, etc.

Conversely, the assessment team may become aware of significant variables affecting outcomes which were entirely unanticipated. Examples include household stability[5], age, birth order, being a non-speaker of the language of instruction in the school due to linguistic mixing, etc. Assessment teams have to be vigilant and inquisitive to identify such variables which may be having an effect on learning outcomes unrelated to the central research question.

4.1.4 Speculative (or Novel). Researchers in any field are constantly seeking to better understand and explain practice and outcomes in their domain of interest. This is true in the field of educational assessment as well. Current models for predicting educational outcomes in low-income countries typically only predict 25-30 percent of the observed variance in performance. Unknown and/or

[4] SocioEconomic Status. In high income countries, the SES variable is often significant in predicting educational achievement. In contrast, in rural areas of low-income countries, it is common to find that there has not yet been sufficient socioeconomic differentiation to impact educational outcomes.

[5] In another MLE experimental program in a low-income country, it was belatedly discovered that more than 50 percent of children were living with extended family members—often grandparents—because of emigration of one or both parents to find work. Local leaders noted that non-parental caretakers were often not so invested in the educational success of their wards.

unspecified variables account for the remaining 70-75 percent of performance. Assessment teams, effectively functioning as researchers, are always interested in identifying novel variables or factors which improve the quality of their assessment models.

Ideally, a good assessment team brings to an assessment situation a good working knowledge of relevant variables which are known to impact or predict educational outcomes. This awareness, along with good skills of observation, curiosity and good judgment allow the team to posit additional educational practices (variables) which may improve learning. Many of these may have been tentatively identified by small-scale studies but benefit from supporting research in a more substantial field study.

4.2 Types of data to be gathered

Literally hundreds of variables have been investigated for their possible impact on educational outcomes. Some, such as large class size, have been consistently found to have a generally negative impact on student learning. Others, such as level of parental education generally have a positive impact on learning outcomes. Many other variables have minimal or ambiguous impacts.

The following sections identify some of the categories of variables which assessment teams normally collect when asked to carry out an assessment of program effectiveness. This listing will not include a lot of detail since the number of variables can get quite large depending on the depth and level of detail pursued by the assessment team. In addition to identifying specific outcomes, each subsection will include some general guidelines and rules of thumb developed from past assessment work.

4.2.1 Learning outcomes. General program assessments almost always include some data on learning outcomes. The options are many: reading skill development, math, science, social studies, writing, learning L2 or L3.

Some guidelines and rules of thumb:

1. Most stakeholders are primarily interested in (a) learning to read, (b) development of math skills, and (c) progress in learning the primary language of instruction in the area. Science becomes important in higher grades but is less so in basic education.
2. Especially in MLE program assessment, all cognitive assessments should be carried out in the language of instruction, not the national language unless that is also the language of instruction.
3. For very young children, a given cognitive assessment should not run longer than 20-30 minutes.
4. Orally administered assessments normally produce more reliable results than written assessments for Grade 2 and below.
5. When paper-based instruments are used, these should use unambiguous response strategies such as multiple choice, matching, doing computations, circling or crossing -out objects, etc. Assessment experience in a number of low-income countries has demonstrated that children quickly master the multiple-choice mechanism and use it with minimal difficulty.
6. In some cultural settings, paper-based instruments may have reliability issues based on cultural patterns or very casual classroom decorum.

4.2.2 Learner variables. It is normal to also gather basic demographic data about the learner including name, age, gender, name of the school, and name of the teacher. Other data is often relevant but may be unavailable or unreliable. This latter includes the following.

1. Attendance record for the school year. Though teachers are generally required to keep attendance records, we have found these to be sketchy, unreliable and often non-existent.

2. Kindergarten attendance. It is widely assumed by most donors that kindergarten gives children a strong head start in basic education. In some countries, this is clearly the case; in others not so much (Walter and Davis, 2003). Nonetheless, such information is important for analytical purposes when doing comparisons in an MLE program assessment. The assessment team must be careful with this one as a **yes** response could mean regularly, sporadically, a few times, etc. If documentary records are available to verify attendance, such data will be more reliable.

It is also true that kindergarten or preschool programs vary dramatically in their quality ranging from being a child-care service to having a full-fledged instructional curriculum.

3. Evidence of malnutrition. Malnourishment has been demonstrated to have significant negative impacts of school performance, but children are not able to answer this question. Teachers or other knowledgeable observers need to supply this data if they have it.

4. Learning disabilities including hearing and vision problems. Younger children will not be able to answer so teachers have to supply this data.

5. The presence of behavioral or attitudinal problems. The teacher needs to supply this data.

6. Ethnicity. Children may or may not know. Teachers usually know but not always.

7. Subjective evaluations. It is also possible to elicit a subjective evaluation from a child's teacher as to the child's educational progress in a skill area of interest. Such judgments, however, may only have a moderate level of accuracy and reliability.

4.2.3 Delivery of Instruction. A wide range of variables fall into this category. Interested parties may recommend variables of interest to them. Gathering reliable data can be costly since the needed data often depends on informed observation by skilled educational evaluators. The following sets of such items is suggestive but not exhaustive for those who may need ideas as to possible data desired.

The teacher. Type, extent, and amount of training; years of experience; proficiency in the language of instruction (LOI); gender; age; level of commitment to education (subjective datum); creativity (subjective); skill in classroom management and organization; etc.

The schedule. Number of subjects in the curricular calendar; instructional hours per day (actual, not official); instructional days per year (actual, not official); average time allocated to each subject per week (actual, not official); mean weekly 'dead time'[6]; teacher absenteeism, etc.

The classroom. Number of students present; adequacy of furnishings; adequacy of physical arrangement of the classroom (lighting, noise levels, visual distractions, temperature), etc.

[6] Dead time includes any of the following: teacher taking attendance, teacher called out of the classroom, emergencies, disorder in the classroom, non-educational discussions or conversations, etc.

Pedagogical practices in the classroom. How instruction takes place; practice time; group work; time spent in review; time allocated to handling student questions, tutoring activities; engagement level of learners in the classroom, etc.

Instructional support. Availability of textbooks and other learning aids, presence of teacher's aides; library resources, etc.

4.2.4 Institutional variables. In low-income countries, it is common to find extreme variation in school effectiveness even among neighboring schools. Institutional variables are much less studied especially in low-income countries with the result that this phenomenon is not well understood. For example, a given school can have an outstanding teacher for one grade and a poor one for the next grade. Another common finding at the level of school administrations in low-income countries is that the headmaster or the principal has no training for the job so that administration is erratic or idiosyncratic. In some cases, the headmaster is very engaged with teachers and teacher development; in other cases, the headmaster may attend only to general administration or local politics assuming that teachers know what they are doing.

Further complicating the gathering of useful data of this type is the general unwillingness of school headmasters to give useful information about their training, their work, or the issues arising from how a given school is run.

Assessment at this level is probably best left to those managing the program. The impact of variation in school effectiveness can be monitored and managed statistically during the analytical process. This will be explained in Section X with focuses on analysis of MLE assessment data.

4.2.5 Contextual variables. Contextual variables can be defined as those describing the physical, cultural, geographic, economic, fiscal, political, or ideological context within which a given school or cluster of schools is located. A common distinction, for example, is the contrast between rural and urban schools. Another is private versus public schools. Some schools might have a tradition of giving special attention to vocational training while others emphasize academic training. Some could be attended primarily by children from a given ethnic community. Others might have an ethnically diverse student body.

The assessment team is likely to find that at least some contextual variables are relevant to their work while others are not or actually point in unexpected directions.

4.3 Specifying the data to be collected: an example from Thailand

The researcher's first task was to define and operationalize the construct of writing skills. Possibilities range from the very basic skill of correctly rendering alphabetic characters to the much more complex task of composing a text. The researcher chose five specific tasks to represent the development of writing skill: letter dictation, writing the initial consonant in a word, writing a word based on a pictorial stimulus, ordering words in a sentence in natural order, and a free-writing task consisting of composing and writing a sentence based on a stimulus word.

Grade 3 children were included in the assessment because these children are receiving literacy instruction in Thai. This allowed the researcher to test hypotheses about the persistence of writing skills developed in Grade 1 as well as applicability to learning to write in a second language.

The statement of the research question provided a direct specification of the data to be collected once the construct of 'writing' had been operationalized. In addition, some basic data on learner characteristics were gathered. Institutional variables were not collected beyond class size since the number of schools involved in the innovation is small. The researcher also gave minimal attention to contextual variables since the make-up of the schools—both experimental and comparison—is culturally and linguistically homogeneous.

The researcher did have to do some post-hoc investigation of learner variables because of an unexpected pattern in the data. This post-doc investigation revealed that three children in one class in one school had rather severe learning disabilities.

5. Step three: Gathering the needed data

Assessment work requires data. Qualitative approaches to assessment require qualitative data—interviews, video and audio records, human observations, field notes, and accumulated program documents—and the application of appropriate qualitative analytical methodology to generate a report (or reports) which sets forth what was learned during the assessment process. Quantitative approaches to assessment also require data but a different type of data. Quantitative assessments are based on quantifiable data such as test results; student records such as age, rates of attendance, dropout rates, graduation rates; surveys; and various types of categorical data which can be quantified for analytical purposes.

The primary focus of this paper will be on quantitative approaches because most stakeholders of assessment are primarily interested in some form of comparative assessment which can be used for decision-making. Accordingly, this section will focus on the gathering of quantitative data for the purposes of assessment.

5.1 Some basic design issues

5.1.1 Identifying the assessment population

Once decisions have been made as to the data to be gathered, plans must be made on how to gather this data. The overall assessment design typically provides a broad definition of the assessment population. At the simplest level, the assessment population is a comprehensive listing of the schools, teachers, and children participating in the intervention—MLE in this case. Usually, however, there is also a second implicit population which serves as a standard against which the innovation is measured. Typically, this is a broader population of schools, teachers, and children participating in the default approach to education. This is often a large population which, in some cases, could be national in scope. However, most assessment projects do not need or want to include the entire national population of schools, teachers, and children in their assessment design. Rather, they will identify constraints which narrow the comparison or baseline population to a manageable level. A common constraint is that of shared ethnicity. In this case, the basic design results in a comparison of educational outcomes among children of a given ethnicity attending experimental schools (those in the MLE program) with similar children attending default—usually—government schools which are using a national language or some other widely used language for instructional purposes.

Geography is another factor which could be used for constraining the comparison population. Others are possible as well though less common.

5.1.2 Sampling strategies

In an assessment design based entirely on written (paper-based) assessment instruments, sampling is less of an issue since administering the instrument to an entire classroom costs little more than administering it to one person or to a small sub-set of the classroom.[7] In a design involving oral or one-on-one assessments, the cost of data gathering can increase dramatically.

When the budget available to carry out an assessment has limited funds (which is normal), the assessment team will need to consider a sampling strategy to keep the assessment within its budget. Sampling may take place at the level of schools as well as at the level of children within a classroom.

The practice of sampling always raises some questions. How does one do sampling? How big should our sample be? What is the potential for error if the assessment project makes use of a sampling strategy? In many contexts, one can also expect to hear the accusation, "This finding may not be valid because of biased sampling."

Since this paper is focused on an innovation in education which spans multiple schools, it will likely be necessary to do multistage sampling if more than 10 or so schools are participating in the innovation. In the first stage, schools are randomly selected from a larger complete list of schools for inclusion in the assessment. In the next stage, x number of children from each school or classroom are randomly selected for inclusion in the assessment.

Whenever random selection is being employed, one must begin with a complete list of schools and then randomly select a subset of these schools for participation. Within each school, a similar complete list of students from the grade or grades in focus will be obtained from which the random sample is drawn.

We cannot, here, provide comprehensive guidance on sampling, but several general principles can be stated:

1. In most assessment situations, random sampling within a defined population is the preferred approach.
2. The potential for error[8] decreases as the sample size gets larger. For example, if the full population of an intervention is 5,000 students, a properly drawn sample of 300-400 will closely represent the entire population when looking at group statistics such as a mean or standard deviation. A smaller sample of 30-40 students has a good chance of deviating significantly from the full population.
3. The greater the number of variables you consider important in understanding the innovation, the greater the sample size needed to support analysis.

Many resources exist on the internet and in print giving guidance and recommendations on sampling. A few of these are given in the list of references and resources.

---

[7] This statement is based only on the cost of gathering the data. Obviously, there will be at least some additional cost involved in scoring instruments and doing data entry. Even these costs, however, can be minimized using technology.

[8] Error, in this case, means getting a result that deviates from the "real" result had the entire population been included in the sample.

5.1.3 Specifying data gathering processes

The specification of data to be collected will be suggestive of the data-gathering processes needed to collect that data. Cognitive assessments will normally require one or more testing procedures. On rare occasions there may already be standardized test results available that one can use though these are almost always administered in the national language which effectively by-passes the effects of the MLE innovation in learning.

When designing an instrument for doing a cognitive assessment, it is common to extend that instrument to also collecting a range of personal data which may impact learning such as age, gender, name of school, kindergarten attendance, educational activity outside of class (such as homework), SES data, details about educational activity in the home environment which support learning, etc.

If data about teachers is to be collected, this usually requires another instrument which is filled out by teachers or by means of an interview with teachers.

Data about classroom practice normally requires direct observation by a trained observer or team of observers.

Information about the school itself may be available from the Ministry of Education or may need to be collected by a visit to the school. If the design calls for data about parental attitudes and involvement in the education of their children, this data will require visits to individual homes and will likely be more difficult to obtain than most of the other information.

5.2 Planning and implementing the assessment design

5.2.1 Developing an approximate timetable

The issue of timing enters into assessment planning at several different levels.

Typically, one must first answer the question, “When should testing of student learning be done?” The answer to this question allows one to begin developing a time frame for the work which must be done prior to any actual testing. The most straight-forward answer to this question is, “Towards the end of the school year.” Much of the educational process is based on “the school year.” At the end of a given school year, teachers and students reach a point in time at which a specified body of material is supposed to have been taught and learned by the students. If a country has standardized tests, this is when they are most likely to be administered. In the case of an experimental program, the end of the school year is a natural point of ‘educational closure’ whether assessment is included or not.

As will become obvious in the next section, substantial attention needs to be given to the time required to develop whatever instruments are going to be required. Planning the logistics, getting any needed permissions, identifying and training data enumerators (testers) all require detailed planning to ensure that assessment work gets completed as intended.

5.2.2 What instruments will be needed and where will these come from? If a proposed assessment project does not have previously developed instruments for the needed testing, these will have to be developed. If an instrument needs to be developed, there is the obvious implication that this instrument will have to be rendered in more than one language. How long will this take? What about piloting a newly developed instrument? How long will that take?

Testing reading in L1. ***A fundamental principle for assessment in MLE program is that basic assessments should be carried out in the primary language of instruction of the child***. In an MLE program, this normally means that reading skill will be tested in one or more minority languages and, in some cases, the national language as well. There is no obvious way one can develop an instrument in multiple languages and guarantee that an instrument rendered in language A provides precisely the same assessment of reading skill development in language B. Fortunately, a well-constructed approximation will produce acceptable results for this purpose.

At least two strategies are available for this purpose. In the first, an assessment team drafts a set of instrument specifications appropriate to the age group(s) to be assessed. With these specifications, a team (former teachers, current teachers, linguists, curriculum developers, ex-teachers) of native speakers can be mobilized to produce a draft instrument in the target language based on the specifications given. This draft can then be reviewed and modified as necessary to produce an instrument to be field tested prior to actual administration. Often, a back-translation of such an instrument to a/the language of wider communication is sufficient for the assessment team to judge the suitability and comparability of instruments when multiple languages are involved.

An alternative strategy is to identify and/or adapt an existing instrument which can be rendered in the target language for this purpose. There are pros and cons to both strategies. The earlier strategy will probably produce a more natural instrument but establishing equivalence in terms of level of difficulty is a challenge. The latter strategy is likely to produce a more stilted instrument in terms of language naturalness, but with great promise of similar difficulty levels across languages and programs. In either case, developers and reviews need a combination of language knowledge, educational awareness, and linguistic awareness to produce appropriate instruments.

Content. The influential National Reading Panel (2000) identified five essential skills of reading: phonemic awareness, phonics, vocabulary, fluency, and comprehension[9]. It is logical, therefore, to develop an assessment strategy which probes the development of these skills—at least for children beyond a certain grade level.

There is an established and available model for such an instrument that the assessment team may want to consider known as EGRA--Early Grade Reading Assessment (Gove and Wetterberg, 2011). This instrument already exists in a wide range of languages. If a suitable version does not exist for your purposes, the team may still find it of value to study how this instrument was designed and how it works so that they can develop their own assessment instrument or approach for reading skill development.

Some caveats. Anyone developing an instrument for testing reading skill development needs to be aware that the development of reading skills in low-income countries typically proceeds at a much slower pace than in high-income countries. Therefore, when testing children in early grades, sub-skills need to be assessed in addition to full-fledged reading comprehension. Similarly, even in higher grades, reading skills will not be nearly as highly developed as is expected in high-income countries. Complex and demanding reading comprehension tasks may well result in zero scores which do not give a reliable indication of progress made in mastering the subskills of reading.

[9] Not all reading theorists have endorsed the findings of the National Reading Panel.

Explicit instruction in phonemic awareness is frequently missing or minimally treated in reading instruction in low-income countries. Furthermore, it is difficult to assess with acceptable levels of reliability. Therefore, the assessment team may want to omit testing in this area.

The basic sub-skills of reading are most reliably tested orally rather than by means of a paper-based instrument. Therefore, for grade 2 children and below, we recommend that reading assessments be done orally rather than by means of a paper-based assessment. Oral assessment usually requires some form of a face-to-face or one-on one testing format. Such testing can be a time-consuming activity which must be taken into consideration in the planning for the assessment (and the training of enumerators[10]).

Testing math skills and knowledge

Testing knowledge and skill in math is typically more straightforward than testing the development of reading skills. In most cases, math assessments can be paper-based and administered corporately to an entire classroom at one time. An exception might be simple assessments done at the kindergarten level or perhaps grade 1 if the learning outcomes are very limited (counting, simple addition involving manipulatives, etc.).

The content of math assessments should be drawn from the national curriculum, from textbooks being used, or from a national statement of learning objectives if the other sources do not exist. The authors' experience has shown that math instruction in early basic education tends to lag the curriculum either because the curriculum is too ambitious (Mereku & Anumel, 2011) or teachers are not well prepared to teach basic math (Akyeampong et al., 2013). Therefore, we recommend that, for grades beyond grade1, at least one third of the content should be drawn from the textbooks or standards specified for the previous year. The other two thirds can come from the objectives established for the current year or the year just completed.

Some more specific recommendations

A common query is, "How many items should be included in a math assessment?" Our experience suggests, "Fewer than you would like to." The following list reflects our experience in this area: Grade 1, 10-12 items; Grade 2, 12-15 items; Grade 3, 15-20 items, Grade 4 and up, 20-30 items depending on the level of difficulty of these. A practical guideline is that an average child should be able to complete the assessment in 20-30 minutes.

At least half of a math assessment for children in Grade 3 and below should be completable without a need to read an instruction or a word problem. This would include tasks such as doing simple calculations (addition and subtraction), filling in a blank, or making a selection from options (some form of multiple choice).

If items are being drawn from the math textbook used in the class, take 80-90 percent of them from the first half of the textbook. The rationale is that teachers very often are unable to cover the entire textbook during the school year.

[10] The term "enumerator" or "data enumerator" is commonly used in some circles in reference to those who do the actual testing and gathering of data.

Current trends in mathematics education tend to favor the teaching of problem-solving strategies versus traditional math facts. Curriculum writers from low-income countries often attempt to introduce the same approach when they return home. Our observation has been that not only do lowly trained teachers not understand this approach; they are often left confused as to what they should teach. For assessment purposes in a low-income country, we recommend that assessments focus more on actual facts and skills rather than some of this more abstract content.

L2 language learning

There is no uniform strategy for assessing competence in a second language. At the highest levels of assessment, L2 proficiency is established by a highly trained specialist who spends at least an hour in conversation with the applicant probing various areas of language competence. At an intermediate level, assessments may consist of some combination of grammar knowledge, vocabulary knowledge, reading comprehension and writing ability using a paper-based instrument. At the lowest level, assessment is typically restricted to reduced subsets of these measures in various combinations.

In the context of an MLE program involving younger children, we have observed strategies as simple as a very brief oral dialogue and as sophisticated as a carefully scaled assessment based on knowledge of increasingly difficult vocabulary items. The former is easy to do (though a bit time intensive) though ultimately subjective. The latter is easy to administer and much more objective but requires substantial work to create the instrument of assessment. Probably the most common is the use of a simple combination of vocabulary knowledge and reading comprehension. This approach is attractive because it is quite simple to construct and administer (either orally or on a written instrument). Since the assessment objective is typically comparison between an innovation and a default approach to instruction, this approach generates useful data for comparative purposes. It does not, however, generate realistic information on actual levels of L2 proficiency, nor a detailed profile of the effectiveness of a given approach to developing L2 proficiency.

Some specific guidelines

For grade 3 and below. Eight to ten vocabulary items which are picturable. These can be divided approximately as follows: four very common nouns, two less common picturable nouns, two common verbs, two common adjectives. Picturing verbs and adjectives often requires a little more creativity and more complex illustrations. The mechanism for testing knowledge can be orally by means of flashcards, or in writing by means of matching, multiple choice selection, or short answer items. Direct matching is the easiest, and providing a short answer is the most difficult as it also requires writing ability.

A short (3-4 sentences) text testing reading comprehension should also be included. The comprehension questions (3-4 questions) should be mostly factual at this level.

For grades 4-6. Children at this level should be able to answer basic grammar questions and understand text material of moderate difficulty. They should also have enough vocabulary knowledge to identify somewhat less common vocabulary including applying the notions of antonyms and synonyms. Assuming a written assessment, the following is a possible mix of assessment items: 3-4 simple grammar questions (pronouns, subject verb agreements, verbal tense markers, noun class markers, etc.); 10-15 vocabulary items using assessment mechanisms such as definitions, pictures, synonyms, antonyms, etc.;

2-3 intermediate length texts (8-12 sentences) with a mix of comprehension questions probing factual detail, inferential understanding, sequencing of events, topic of the story, etc.

If the assessment question requires probing of productive ability in L2 (writing, speaking), these can be included as well. Obviously, an assessment of speaking ability will require some form of one-on-one interaction probably with a fixed set of questions to be answered.

Other subjects such as science, social studies, environment, religion, etc.

Data gathered in multiple countries indicate that the amount of time given to these subjects as well as the content of instruction can vary substantially in low-income countries (World Bank, 2017). Furthermore, some teachers adhere closely to curriculum guides while others may largely ignore them especially in the lower grades.

Accordingly, we only recommend the inclusion of content on these subjects when key stakeholders require that it be included. Even then, coverage can usually be minimal.

Guidelines and recommendations

1. In general, we recommend not including assessment in these areas below grade 4.
2. Use the multiple-choice mechanism for testing in this area.
3. Keep questions as non-controversial as possible so that student responses are easier to interpret.
4. Keep the number of questions as limited as possible.

Other instruments

If the planned assessment is to include substantial detail on teachers, classroom practices, family literacy practices, etc., instruments will need to be developed for these purposes. Because the content of such instruments is apt to be varied from one assessment to another, we are not able to provide much specific guidance beyond some general principles of instrument construction.

Guidelines and recommendations

1. Keep instruments focused; avoid collecting data which has not been planned.
2. Many of the questions that one might want to ask of teachers are scalar in nature and require that teachers answer by means of a 5 or 7 point scale (e.g., Always, Usually, Not sure, Occasionally, Never). This type of a scale is (a) hard for people to respond to without considerable experience, (b) prone to the respondent giving 'proper' or 'expected' answers, and (c) and vulnerable to a polar response pattern (answering either Always or Never).
3. Many teachers and school administrators will want to protect themselves and their positions by how they respond to questions posed by an outside assessor. Consider seeking to develop questions which generate desired data somewhat indirectly. This understandable tendency on the part of education personnel is one reason why data gathering in this area is often done via direct observation or via video. Both of these methods are time-consuming, costly, and often difficult to evaluate.

What about piloting instruments before the official assessment begins?

Piloting instruments is always a good idea but one does not always have the time or the resources to do a good pilot. What does a pilot reveal about an instrument? The following is a list of features about an instrument that can be learned from a pilot test of that instrument.

1. A primary object of piloting an instrument is to identify any issues in item construction or presentation. A second is to gauge the amount of time needed to complete the instrument.
2. A primary objective for piloting an oral instrument is to give enumerators practice in doing the administration.
3. Extremely low performance. If children are entirely unable to complete the instrument being piloted or effectively get a very low score, the items on the instrument are probably too difficult. It is also possible that children have not learned any of the content. If all children get a low score, the test items are probably too difficult or there is some other issue in teaching or learning. Assessment teams should be aware that when younger children are being taught in a second language, unusually low scores are probable even though the content is correctly pitched for the grade level of the children.
4. Extremely high performance. If most or all children complete the instrument quickly and/or get very high scores, the items on the instrument are likely too easy. If the content is such that it is quite certain that not everyone should get a high score, then the instrument should be adjusted accordingly.
5. Everyone answers the same item incorrectly and in the same way. This pattern indicates that the item in question has been poorly written or is miss-leading in some way. The item can be reviewed and repaired or replaced.
6. An item is confusing and generates questions or puzzled looks. Such an item is probably poorly worded or poorly constructed. Occasionally, if an item is multiple choice, the correct answer may not have been included as one of the options.
7. A well-designed assessment of mastery of specified content should generate a mean score of between 40 and 60 percent. A mean score below this indicates that content has not been well-learned or that too much of the content is too difficult. Mean scores above 60 percent generally indicate a high level of mastery or content which is pitched below the skill level of students.
8. When planning a pilot, common questions include, “How many students should be included in the pilot,” and, “Can we include children in the pilot who will be included in the regular assessment?” The answer to the first question is that, under most circumstances, it is not necessary to include a large number of students in a pilot to get the needed information on suitability. Fifteen to twenty students are usually sufficient for this purpose.
9. On the second question, we have normally made it a practice not to include children in a pilot who are going to participate in the regular assessment. However, there are caveats. When children are younger, they will not benefit or learn much from a pilot, so little distortion is introduced into the full assessment. A bigger issue is external perception of how the pilot was done. Observers, especially critics, will see such a pilot as biasing the results in favor of either the experimental group or the comparison group. For this reason, we have normally chosen not to include anyone in a pilot who is to participate in the full assessment. In some situations, this may mean that some creativity or judgment is needed as to how to test the suitability of an instrument.

5.2.3 Selecting and training enumerators

The number and type of enumerators needed will obviously depend on the scope and design of the assessment. Again, there are too many variables to cover every possible combination or contingency, so we include only some hints and guidelines.

Some guidelines and recommendations

1. Oral assessment such as EGRA is best done with a team of two people working together. One interacts with the student or learner and one records the student's performance..
2. The data enumerators need to speak well the language of the children. This is especially important for those in an MLE innovation as children may not understand anything spoken in the national language or language of wider communication.
3. It is very helpful if the enumerators have had past experience in basic education such as retired or former teachers.
4. If oral assessment is part of the design, the enumerators not only need training but also practice in the administration of the assessment.
5. Enumerators need to be familiar with the geographic area where the assessments will be carried out especially when schools are in rural areas and not always close to good roads.
6. Careful thought has to be given to the number of enumerators that will be needed to complete the assessment work. The number will depend on the type of assessment, the amount of travel which needs to be done, the length of the school day, the accessibility of the schools, and the budget resources available to pay for their work.
7. Inevitably, one finds that some data enumerators are more careful, thorough and conscientious about their work than others. The assessment team needs to take steps to monitor the work of their data enumerators and make decisions about assessment work (data) and workers which do not meet the standards set for the assessment.

5.2.4 Management of the data-gathering process

As in the section above, it is not possible to give detailed guidelines for every imaginable assessment scenario one might encounter. Therefore, we limit our comments to significant guidelines and reminders that will generally apply to all or most assessment projects.

Some guidelines and recommendations

1. One or more fulltime staff need to participate in the day-to-day management of the data-gathering process.
2. All managers as well as data enumerators need to have one or more phones available to them at all times to deal with the issues encountered.
3. Managers should keep a log of problems encountered and solutions agreed to so that disagreements can be resolved by means of a written record rather than the memories of those involved in the process.
4. Managers need to communicate plans for the coming day of work at least a day ahead of time to allow time for those in the field to check their supplies (for the next day), make local travel arrangements, seek any local permissions needed, and make a schedule for the day which can be communicated back to the manager ahead of departure into the field.
5. Managers need to coordinate with local school[11] and government officials to ensure that everything is in place for the coming assessment work.

[11] The issue of whether to alert teachers or headmasters to the fact that data enumerators are arriving on a given day to do assessment is something which needs to be discussed with local officials as to whether this step helps or hurts the integrity of the assessment.

6. Managers need to have contingency plans in place for emergencies such as storms, sickness, accidents, vehicle breakdowns, team breakdowns, school closings, and other unanticipated interruptions in the data-gathering process.

5.2.5 Scoring instruments

Before scoring begins, it is a good idea to write on every copy of each instrument the name of the school, the grade from which the data came, the month and year in which the data were collected, and the name(s) of those who collected the data. It is also a good idea to have large envelopes or folders available into which to put all of the instruments of the same type from a given school and grade. Inevitably, there will need to be some error checking for missing data or data incorrectly entered into the database.

The following are guidelines and recommendations on the process of scoring instruments, especially cognitive assessments.

Written assessments

1. Scoring templates should be created for marking written cognitive assessments.
2. To the extent possible, all items should be either correct or incorrect. Giving partial credit only complicates data entry and adds to the level of subjectivity present in an assessment. What constitutes a correct answer should be established before scoring begins.
3. It is a good idea to involve at least two people in scoring. Scorers can work in pairs with one reading answers and the other marking the instrument.
4. There is no need to compute totals or averages by hand as this is time-consuming, error prone and something which can be done very quickly with software.

Oral assessments

Scoring guidelines need to be carefully thought through and agreed upon before data collection begins. For example, many orally administered instruments are also timed so that those being tested are tempted to go too quickly or to do a lot of guessing. Therefore, decisions have to be made about the following behaviors:

1. If a student skips items, should those be marked as incorrect, or not attempted (like missing data)?
2. What about a quick self-correction after an initial incorrect response? Should such instances be marked as correct or incorrect?
3. What if a student is obviously guessing and one of the guesses is correct? How is that marked?
4. How much time should a student take trying to formulate a response before being urged to move on to the next item?
5. What if the student gives an initial correct response and then “self-corrects” to an incorrect response? How is that behavior scored?
6. The guidelines for scoring oral instruments need to tell scorers how to mark each of these behaviors. There is not necessarily a right or wrong way to mark these behaviors; the important point is that all enumerators follow the same standards in marking such responses.
7. Oral assessments such as the EGRA include tests of a skill having many like elements such as identifying letters of the alphabet, or reading lists of words. When such a task is included in an oral assessment, it is usually helpful for the scorer to go ahead and count all the correct responses and all of the incorrect responses to facilitate data entry since the data entry will probably be shortened just to basic totals rather than recording performance on each element of a subtask.

5.2.6 Data entry

In most cases, entering data into an EXCEL spreadsheet or worksheet will be the most feasible option for initial data entry. In general, we recommend that a person who is going to be involved in analysis create an EXCEL template into which the data is to be entered. It is normally a good idea that the creator of the EXCEL worksheet also create a codebook for the worksheet when it is completed. The codebook explains what each variable name (column label) means and includes a definition or explanation of how data is to be entered into a given column. EXCEL allows a person to enter multiple types of data (e.g., a word, a number, a percent, a comment, etc.) into the same column. Virtually all statistical software requires that the data in a given column all be of exactly the same type. Mixing of words and numbers is strictly forbidden.

The standard data structure for most statistical software is variables in the columns and cases in the rows. For an MLE assessment, this will mean that student names or an ID number will be in the first column as each student is a "case" or "record" in the jargon of database structures. Other information about data entry is found in the following list of guidelines and recommendations.

Guidelines and recommendations for data entry

1. In our experience data entry works best when done in pairs with one person reading the data and another person doing the keyboarding. In this way, both are able to look at the monitor to assure that data has been correctly entered.
2. If there is a need to do data entry simultaneously on multiple computers, this can be done AS LONG AS the same data template is being used. When this is done, it is absolute crucial that no columns be added to or deleted from either of the worksheets. Doing so readily leads to DISASTER when the two worksheets are merged into one for analytical purposes.
3. Data entry can be a rather monotonous and robotic process so it is a good idea for the keyboardist to take frequent breaks or to switch places with the reader in order to stay fresh.
4. There are several data entry conventions which need to be observed to maintain accuracy and fidelity in the assessment. A very basic convention is that a 1 indicates yes or a correct answer and 0 (zero, not a capital o) indicates no or an incorrect answer to a question. Entering a lower-case L for a 1 or an uppercase letter o for a zero will create problems at the analytical level. If a keyboardist has a tendency to do data entry in this way, that person should be reminded of the importance of using numbers and not letters which look like letters.
5. Another common issue in data entry is that of "missing data." If a given variable (column) requires a word or string of letters, leaving a cell blank in that column means the needed datum is not available. However, in the case of a column requiring a number (especially one requiring a 1 or 0 to indicate correct or incorrect, a blank cell may mean that there was no response for a given test item OR, it can mean that the keyboardist came to a section which was beyond the ability of the student and decided to leave it blank rather than take the time to enter a zero. The assessment management team needs to decide before data entry begins how to handle incomplete responses. That is, does a non-response on a test indicate only that the item was inadvertently overlooked by the test taker and might have been answered correctly if not overlooked, or does a non-response indicate that the test-taker was unable to give a response because of a lack of knowledge or skill. Because of the nature of the assessment (measurable evidence of knowledge and skill gained), we recommend that non-responses on a cognitive assessment be treated as an incorrect response with a zero entered into the data sheet for that person for that question.[12]

[12] The rationale for this recommendation is the following. It is highly likely that a person who can answer 60 percent of all test items correctly has learned more than a person who responded to only one test item and

6. When doing data entry, data should be saved frequently and on multiple devices to avoid irretrievable loss of work completed.
7. Once data entry has been completed, the data sheets should be checked for the possible entry of spurious numbers such as 11 or 12 rather than 1 for a correct response or a 9 rather than a 0 for incorrect responses. It is also good to check such columns for the presence of lowercase L's and uppercase letter o's.

5.3 Gathering the needed data: an example from Thailand

The researcher decided to collect data from two grades in two experimental and two comparison schools. All children in the two grades in the four schools were included so no sampling was needed.

The data gathered were both quantitative and qualitative (classroom observations). Custom instruments were developed for the assessment since none existed. These were developed in both Thai and Malay so that students could respond according to the language of instruction. The instruments were piloted with children in the following grade to identify any issues in design. The researcher trained data enumerators from a local university to administer the cognitive instruments but gather the qualitative data herself.

All of the scoring was done by the researcher using rubrics especially designed for this purpose. The researcher also did all of the data entry since the data set was relatively small.

6 Data analysis

Analysis of the data is the most technically demanding part of an assessment project in MLE. Realistically, the analysis should be done by a person or a team with some training and experience doing quantitative data analysis. If the assessment team does not include someone with the requisite expertise to do the task, then someone needs to be found or hired to do the work. The details, processes, decisions, and methodologies involved cannot be adequately spelled out in a short paper to enable a person without a suitable background to do the needed analysis. However, we can provide enough of an overview of the task to help you communicate with a specialist if you need to bring one in for this purpose.

6.1 Phase 1 – Data cleansing

All or virtually all imported data sets have to be cleansed to prepare for analysis. Data cleansing includes but is not limited to the following data checks.

6.1.1 Checking to ensure that numerical values fall within range.

1. Checking categorical variables for spelling errors and other inconsistencies (which are very common). For example, it is common during data entry to enter both upper and lower case values of f and F indicating female and m and M indicating male. Most software will treat these as indicating four separate values for the Gender (or Sex) variable.

answered that item correctly either by random chance or by having learned just that one small piece of information. If non-responses are discarded then most analyses will indicate that the first person scored 60 percent correct while the second person scored 100 percent correct. That is, we prefer an assessment which more highly values breadth of learning over the ability to correctly answer a very limited number of questions correctly with a high level of certainty.

2. Checking for missing data and deciding what needs to be done (if anything).
3. Checking for instances of the wrong data type entered into a column (e.g. a number rather than a word, or a decimal rather than a percentage).
4. Checking for the presence of space or other characters in data columns which can lead to significant and sometimes mystifying errors in the analysis.
5. Checking for implausible values caused by miss-typing (e.g. an age of 61 for a child, a gender identifier as FF, a final score of 150 percent on a test, a distance value of 200km when 20km was intended, etc.
6. Checking for accuracy when synthetic values such as test totals are created automatically in EXCEL upon data entry.

6.2 Phase 2 – Computing synthetic variables

Before analysis begins, it is a good idea to save a master copy of the 'cleansed' data set in multiple locations. Doing so makes it possible to replace data which may be accidentally altered or deleted in the process of sorting the data, creating synthetic variables, and moving data around for whatever purpose.

A normal early step in the analysis process is to create a range of synthetic variables[13] to facilitate analysis. Often, synthetic variables are needed because they provide a more meaningful summary of findings than the raw data itself.

In other cases, synthetic variables are needed just to support analytic activity. A very common example is the use of so-called "dummy variables". At the time of data-gathering, gender is typically indicated by a letter or word such as "F" or "Female". Some analytic processes are not able to use data values in the form of letters. Rather, letter values such as those for gender have to be converted to numbers (such as "F" = 1, "M" = 0 or the reverse depending on arbitrary choice). These numbers can then be used for computational purposes. The meaning is not lost but transformed in a way that software can understand.

6.3 Phase 3 – Computing summary statistics and testing basic hypotheses

Computing summary statistics—mean, standard deviation, N—is a good first step for developing an initial understanding of what the data says about the innovation in focus. Besides helping the analyst develop an idea of how to proceed with analysis, having such basic summary statistics helps satisfy the intense curiosity of the stakeholders invested in the program.

These summary statistics should be broken down by the major categories described or implied in the design—program models, grades, geographic regions, etc. In many cases, the analyst will also do basic hypothesis testing to determine the effect size of the intervention driving the assessment. Considerable caution needs to be exercised, however, in how much information or how much detail to reveal early in

[13] A synthetic variable is a new variable created by the analyst from raw data present in the worksheet to express important information derived from the data. For example, a written test may be composed of 4 sections each of which consists of anywhere from 3 to 8 test items. The analyst knows that much of the analysis will revolve around overall performance on this test so a composite score—usually a percentage of items correct—is needed for the entire test. The overall test score—expressed as a percentage—is a synthetic variable because it is dependent upon and derived from other original variables in the data set.

the analysis. This is for two reasons: first, stakeholders who have major or controversial agendas may misuse preliminary findings in promoting their agendas, and second, as analysis goes deeper, findings may shift in significant ways in terms of effect sizes and the impact of various independent variables on the early findings.[14] Our recommendation is to be very cautious in what preliminary findings are released and to whom.

6.4 Phase 4 – Testing basic explanatory models

Most assessments will collect some information about individual participants in the review. These variables become ones which the researcher may want to analyze more deeply to determine their impact on learning. Variables such as age, gender, kindergarten attendance, levels of absenteeism, SES level are all known to impact learning in basic education. These variables can be probed singly by means of appropriate tests or collectively by means of strategies such as multiple regression.

Ultimately, the goal is to construct a model which best accounts for individual student learning while simultaneously maximizing predictive power and minimizing the number of variables included in the model. There is no magic formula for settling upon an optimal model so one has considerable freedom to explore various combinations of variables and models.

Some of the variables included in the design apply most directly to schools rather than individuals though individuals are the normal focus of measurement. Investigating models which explain performance differences between individual schools such as location, class size, teacher experience, etc., may also contribute to an improved understanding of the MLE program being assessed.

6.5 Phase 5 – Testing more advanced explanatory models

It is widely recognized that schools vary in the quality of instruction they provide. An implication is that every child in the school is impacted in the same way by such variations in quality of instruction. Only in recent years have analytical strategies been developed for simultaneously modeling individual-level and school-level variables. This approach is variously referred to as multilevel modeling or hierarchical linear modeling. The use of multilevel modeling allows the researcher to control for school-level effects while seeking to interpret the impact of an intervention on the performance of children.

Multilevel modeling adds considerable conceptual complexity both to the completion of the analysis and presentation of results to stakeholders who often do not understand this approach to analysis. If these more advanced models are included in your analysis of results, you will need to report findings in a way that stakeholders can understand.

6.6 Data analysis: An example from Thailand

The research question entailed an explicit hypothesis: learning to write in Malay (the L1) would facilitate learning to write Thai (the L2). In the process of operationalizing the construct of “writing”, the research

[14] A good example of this is a study of private schools in Latin America (Somers et al, 2004). For many years, testing evidence showed that private schools significantly out-performed public schools in the quality of education provided. This enabled private schools to attract more students and funding. However, when researchers controlled for the individual and classroom-level socioeconomic background of students (SES), it was discovered that private schools did not provide superior outcomes. Rather, an apparent advantage appeared to be due to the students they attracted, not the quality of instruction they provided.

articulated additional and more specific hypotheses regarding differential abilities in (a) letter formation, (b) writing a word-initial consonant, (c) ability to specify correct word order in a target sentence, (d) ability to correctly render a vocabulary item, and (e) ability to construct and render a simple sentence. In all five skills tested, the children in the experimental program performed at a higher level than children in the comparison schools. In only one case—correct word order—was the difference not statistically significant.

The grade 3 task required students to do free writing on a topic of their choice for 30 minutes. These student products were then evaluated using a three-part rubric testing (a) mastery of language features, (b) assessment of content and (c) assessment of organization and coherence. Directional hypotheses were constructed and testing in keeping with the overall hypothesis that learning to writing in L1 enhanced one's ability to write in L2. On this assessment, the experimental children outperformed the comparison children on each measure, but the difference was statistically significant only on the measure of language features. Given the relatively small sample size, the research did not pursue the testing of more abstract models.

7 Reporting results

The preparation of one or more reports of the findings from the assessment brings another set of challenges for the assessment team. What are the information needs of the various stakeholders? How much detail should be provided? How do you deal with the various levels of ability to understand complex detail in a report? In what language or languages should the report be written (or translated into)? How long should the report be? Should the assessment report be made public? Who "owns" the assessment report? How should the report be used and for what purposes? Should there be multiple versions of the report for different stakeholders? How, when, and too whom should be report be presented?

There are no easy answers to some of these questions. Typically, the primary report should be written for the primary stakeholder whether that be the donor, the ministry of education, or some other entity. Normally, however, there are multiple stakeholders so the needs of each will have to be taken into consideration.

Rather than attempting to anticipate a large number of scenarios and what constraints each brings to the need to write and/or adapt reports, we will present some general or broad guidelines to help think through the matter of reporting the findings of an assessment.

Guidelines and recommendations

1. A substantial report is usually required to adequately cover the details of the assessment and the findings. Most readers, however, are primarily interested in the central or most important findings. Therefore, even when writing a lengthy report, we strongly recommend writing a short executive summary of 2-4 pages which presents a response to the primary research question and highlights any other critical or unexpected findings from the assessment. Many key stakeholders also like to see a section on recommendations if the team has any.
2. Governments are primarily interested in implications for policy making and program implementation.
3. Donors are primarily interested in cost-effectiveness and implications for the allocation of resources in other similar settings.

4. The academic community is interested in findings which impact current theory on multilingual education. Surprising or unexpected findings are of special interest in this regard. The academic community will always be on the lookout for effects coming from unexpected sources.
5. Implementing agencies are interested in issues such as program effectiveness, areas of strength and weakness, and steps which can be taken to improve program delivery.
6. Media outlets (which may or may not come into the picture) are interested in sound bites on topics of significant current interest including, in some cases, matters of controversy.
7. We recommend that the assessment team have some early discussion with major stakeholders about the kind of report(s) they would like to see. This is a good strategy to help ensure that attention is given to topics which concern them and a good way to identify potential "land mines" to avoid.
8. If the initial version of the report is written in a language other than the working language of the country in which the innovation is taking place, it will be appropriate to discuss the need for some version of the report in that working language.
9. Many readers will have a minimal understanding of statistics (or none at all) so textual explanations or graphs are also needed for clarity of communication.
10. Reducing technical findings to simple percentages showing the difference between two programs or the percentage of improvement brought about by the innovation is often appreciated.

7.1 Reporting results: An example from Thailand

In the example being reported in this paper, the results were reported in the researcher's dissertation for the doctoral degree. Less formal reports were made to the local community and to government officials interested in the experimental model of MLE being tested in the Malay community.

Given the level of formality that a dissertation represents, it would be entirely appropriate to publish less formal reports to other sectors of the Thai professional community who might be interested in the findings.

8. Conclusion

In this paper we have sought to provide a reasonably detailed guide to planning and implementing an assessment of an experimental program in multilingual education with a focus on low-income countries.

Carrying out such an assessment is a major project which is one reason why such assessments are relatively rare outside of major development agencies such as the World Bank or US AID. Even if the management of a given experimental program is not planning to do such an assessment internally, the background provided should be helpful in working with a contractor to carry out such a project.

The list of references includes some suggestions of possible resources one might consult for ideas on the implementation of various components of an assessment study.

References and Resources

Akyeampong, K., K. Lussier, J. Pryor, & S. Westbrook. 2013. Improving teaching and learning of basic maths and reading in Africa: Does teach preparation count? *International Journal of Educational Development,* 33:272-282.

Burarungrot, Mirinda. 2016. *Writing skills transfer from Patani Malay (L1) to Thai (l2) in mother tongue-based bi/multilingual education schools: a case study in Patani Malay-Thai in southernmost provinces of Thailand*. PhD dissertation presented to Mahidol University.

Chingos, Matthew M. 2012. Strength in Numbers: State spending of K-12 Assessment Systems. Washington D.C., Brown Center on Educational Policy at Brookings.

Davis, Patricia. 2004. *Reading is for Knowing: Literacy Acquisition, Retention and Usage among the Machiguenga*. Dallas, TX.: SIL Int.

Dunn, Douglas M. *The Peabody Picture Vocabulary Test-V5*.

https://www.pearsonassessments.com/store/usassessments/en/Store/Professional-Assessments/Academic-Learning/Brief/Peabody-Picture-Vocabulary-Test-%7C-Fifth-Edition/p/100001984.html Last accessed 01/28/2021

Eunice Kennedy Shriver National Institute of Child Health and Human Development, NIH, DHHS. (2000). Report of the National Reading Panel: Teaching Children to Read: Reports of the Subgroups (00-4754). Washington, DC: U.S. Government Printing Office. https://www.nichd.nih.gov/publications/product/247 Last accessed May 6, 2021.

Gove, Amber and Wetterberg, Anna. 2011. *The Early Grade Reading Assessment: applications and interventions to improve basic literacy*. Research Triangle Park, NC.: RTI Press.

Laitin, David D., Ramachandran, Rajesh, and Walter, Stephen L. The legacy of colonial language policies and their impact on student learning: evidence from an experimental program in Cameroon. *Economic Development and Cultural Change*, 68(1) 239-272.

Mereku, D.K. & C.R. Anumel. 2011. Ghana's achievement in mathematics in TIMSS 2007. *Mathematics Connection*, 10, 83-99.

Mertens, Donna M. and Wilson, Amy T. 2019. *Program Evaluation Theory and Practice, 2nd ed.* New York, NY.: The Guilford Press.

Osterlind, Steven J. 1998. *Constructing Test Items: multiple-choice, constructed-response, performance, and other formats*. Boston, MA: Kluwer Academic Publishers.

Somers, Marie-Andree, Patrick J. McEwan and J. Douglas Willms. 2004. How Effective Are Private Schools in Latin America? *Comparative Education Review*. 48(1), 48-69.

Walter, Stephen L. 2016. *The EMBLI Endline Evaluation Study*. Research report presented to the Ministry of Education in East Timor, pp 1-193.

Walter, Stephen L. and Patricia M. Davis. 2003. *Eritrea National Reading Survey: 2001-2002*. Dallas, TX.: SIL Int.

World Bank. 2017. *Lessons Learned from an Early Assessment (2017) of Two Innovations in Basic Education in Timor-Leste*.