
# WHITE PAPER SERIES

## Centers for Excellence in Teacher Training (CETT) Program

## Paper Two: Testing and Assessment

**FINAL REPORT**

January 2012

This final report was prepared for the United States Agency for International Development (USAID), Bureau for Latin America and the Caribbean, Office of Regional Sustainable Development, Education and Human Resources Team, under the Evaluation and Technical Support to LAC/RSD/EHR Contract AFP-M-00-06-00047-00. It was prepared by the Aguirre Division of JBS International, Inc. Its primary authors are Gilbert Valverde, Richard Wolfe, and Renzo Roncagliolo.

## CETT WHITE PAPER SERIES

This document is one in a series of white papers discussing the implementation and outcomes of the Centers for Excellence in Teacher Training (CETT) program. The CETT program was implemented by USAID's Bureau for Latin America and the Caribbean, Office of Regional Sustainable Development, Education and Human Resources Team from 2002–2009. CETT was based on a Presidential Initiative derived from commitments made by the U.S. Government at the Summit of the Americas in 2001 and operated in twenty-one countries in the regions of Central and South America, as well as the Caribbean.

The purpose of this CETT white paper series is to highlight the legacy of the initiative and to provide future program designers with some of the most important lessons learned and best practices developed within the long-term implementation of the CETT program.

The CETT white paper series includes five publications by theme:

**Paper One: Regional Nature**
*This white paper discusses the challenges, successes, and lessons learned implementing a regional model for teacher training. The regional nature of CETT differentiated this program from other, strictly national, teacher professional development efforts undertaken by USAID. Three CETTs in the Caribbean, Central and South America underwent a significant process of compromise and cooperation to arrive at their regional models and this paper documents the initiatives taken.*

**Paper Two: Testing and Assessment**
*This white paper discusses the challenges and lessons learned in the process of creating a cross-country testing initiative. The three CETTs carried out testing initiatives to track student performance toward literacy benchmarks, with the goal of showing valid and reliable results. An extremely challenging endeavor, student assessment is further complicated when using tests across countries.*

**Paper Three: Sustainability**
*This white paper discusses the lessons learned while anticipating the challenges of sustaining the CETT program after the end of USAID funding. The CETTs worked closely with USAID to prepare for the continuation of the program at the regional, national, and local levels. The paper examines the political, financial, institutional, and social sustainability dimensions of these efforts.*

**Paper Four: Paradigm Shift**
*This white paper discusses the systemic change in the behaviors and attitudes of CETT stakeholder groups, including school administrators, teacher trainers, teachers, parents, and students. CETT's teacher training model stressed the inclusion of stakeholders at all levels to promote the importance of reading and writing. Achievement of the program's intended effects depended on the willingness of the institutions and individuals involved to change their behaviors. This paper highlights the lessons learned and best practices in promoting this change.*

**Paper Five: Cost Effectiveness**
*This white paper presents a cost-effectiveness study linking financial inputs and CETT program outcomes. The CETT model of teacher training developed differently in each of the three regions and this white paper analyzes the history of costs over time, cost-effectiveness based on teacher and student performance, and the limitations of comparing costs across countries and programs.*

# WHITE PAPER SERIES

## Centers for Excellence in Teacher Training (CETT) Program

## **Paper Two: Testing and Assessment**

Prepared by:

Gilbert Valverde
Richard Wolfe
Renzo Roncagliolo

Edited by:

Aguirre Division of JBS International, Inc.

# TABLE OF CONTENTS

## Acronyms

| | |
|---|---|
| CA-RD | Central America and Dominican Republic |
| CARICOM | Caribbean Community |
| C-CETT | Caribbean CETT |
| CETT | Centers for Excellence in Teacher Training |
| CRSAT | Caribbean Reading Standards Achievement Test |
| ICT | Information and Communication Technologies |
| IRT | Item Response Theory |
| LAC | Latin America and the Caribbean |
| USAID | United States Agency for International Development |

## Introduction and Methodology

The Centers for Excellence in Teacher Training (CETT) program was a Presidential Initiative to improve the pedagogical skills of teachers in the first, second, and third grades in economically disadvantaged communities of Latin America and the Caribbean (LAC). The hemisphere-wide program—announced in 2001 and implemented by the U.S. Agency for International Development (USAID)—created three regional[1] CETTs that began implementation in 13 countries, referred to in this study as:

1. C-CETT (beginning in the Caribbean countries of Jamaica, St. Vincent and the Grenadines, St. Lucia, Guyana, and Belize);[2]
2. Centro Andino (Ecuador, Peru, and Bolivia in South America); and
3. CETT CA-RD (in the Central American countries of El Salvador, Guatemala, Honduras, and Nicaragua; and in the Dominican Republic).

The Cooperative Agreements for USAID assistance to the CETT program ended in December 2009 after over seven years of technical support. (Two CETTs were issued a no-cost extension until early 2010.) As a result of the program, 35,095 teachers and administrators received training in interactive methods of literacy instruction. The program reached over 799,000 students in 21 countries.

CETT provided in-service training to teachers and administrators located in disadvantaged rural and urban areas that did not benefit from other donor programming. The program promoted the development of skills and adoption of active-learning strategies for teaching reading by aligning existing pedagogical practice with research-based best practices. The program had five core components:

1. **Teacher training** in effective reading methodologies and classroom management techniques
2. **Materials** for teachers to use to improve their reading instruction
3. **Diagnostic tools** to enable teachers to identify and address students' weaknesses and needs
4. **Applied research** to ensure the efficacy of the training, tools, and materials provided
5. **Information and communications technologies (ICTs)** to broaden access to the program

In addition, the CETTs also focused on sustainability efforts to ensure continuance of the program after the end of USAID funding. Within the parameters of these components, each CETT had the flexibility to manage and implement the program based on its regional context and needs. As a result, the CETTs developed with slight differences in each region.

CETT training *content* was related to seven literacy skills: reading comprehension, phonological awareness, phonics, fluency, oral expression, written expression, and vocabulary. Knowledge of these skills provided the foundation for integrated and effective reading instruction.

[1] In this study, "regional" refers to one of the three CETT areas: the Caribbean, South America, or Central America and the Dominican Republic. "Hemispheric" refers to all three CETTs as a single unit.

[2] By the end of the program in 2009, many more islands in the Caribbean had adopted CETT. Jamaica, St. Lucia, St. Vincent and the Grenadines, Belize, Guyana, and the Commonwealth of Dominica implemented CETT with USAID funding. After learning of the experiences and results of other countries, the governments of Trinidad and Tobago and Grenada approached C-CETT to join, fully financing their own implementation and purchasing C-CETT's technical support. In 2009, five additional countries signed Memoranda of Understanding (MOUs) to expand CETT implementation to St. Kitts and Nevis, Antigua and Barbuda, Anguilla, Montserrat, and the British Virgin Islands.

The CETT teacher training model introduced innovative *techniques* such as continuous teacher training throughout the school year and follow-up support in the classroom. Teacher trainers visited CETT classrooms where they observed teachers and provided feedback and recommendations. Teacher circles gave teachers the opportunity to share their experiences with peers. Each CETT also emphasized the role of parents and the greater community in embracing a "culture of literacy" to support the importance of reading in the early grades.

The program was implemented in two phases: Phase One (2002–2006) and Phase Two (2006–2009). Phase One launched the initial CETT program design and development. Lead implementing institutions in Jamaica, Honduras, and Peru signed Cooperative Agreements with USAID. Phase Two supported a continuation of the CETTs following USAID's emergent consensus that five years were not sufficient to fully implement the program and achieve the desired results.

## Purpose

This white paper examines the experience of CETT in designing and developing a student assessment system to measure program impact – i.e., using student achievement tests to measure progress at the program level. It focuses on the student testing component of CETT's evaluation strategy only. It looks at the challenges faced and the lessons learned from the cross-country student testing effort. The CETT experience is compared also with international best practices in student testing for program evaluation in order to draw recommendations for similar program initiatives in the future.

The first section of the paper provides a general description of program evaluation in CETT, which encompassed a number of activities, including the testing initiative that is the focus of this paper. The second section presents the three fundamental challenges to the evaluation design given the monitoring and evaluation efforts of the CETTs, and how these challenges affected the testing initiative. The third section documents the lessons learned and the innovative techniques that the CETTs used in designing the student tests, and in analyzing the test results. In the final section, the research team proposes a framework for action in the form of recommendations based on the CETT experience that can inform testing initiatives within program evaluation of future USAID interventions.

This study is part of the CETT white paper series, a compilation of five research papers on key topics related to CETT: regional nature, testing and assessment, sustainability, paradigm shift, and cost effectiveness. Each of the white papers examines the three CETTs through a selection of lenses and analyzes the research findings to bring significant and specific lessons learned with respect to CETT activities into focus. This research gives form to the legacy of the Presidential Initiative and provides future program designers with some of the most important lessons learned during the long-term implementation of the CETT program.

## Research Questions

The research hypothesis of this white paper is that the use of evaluation and continuous assessment helped develop practices and improve program outcomes while building local capacity in the area of monitoring and evaluation. The research team, led by expert consultants Dr. Gilbert Valverde and Dr. Richard Wolfe, reflected on the thoughts, responses, and attitudes of CETT program staff from all three regions in order to examine both what went well and what was overlooked in CETT's testing initiative. This study set out to investigate several research questions:

1. For each region and overall, to what extent was the testing initiative successful in building regional testing capacity and assessments that accurately show outcomes resulting from the CETT interventions?
2. What were the advantages and synergies of the regional approach to the CETT testing effort that are noteworthy for future programming? What were the limitations in the process of developing these?
3. As each CETT developed its testing approach and methodology, what differences developed among them? What strengths and weaknesses did each approach have? How could they have been improved?
4. In what ways/to what extent did the CETTs meet the challenge of fielding evaluations adhering to acceptable standards of accuracy, feasibility, and utility? To what extent did the testing systems established by each CETT contribute to a meaningful and valid evaluation of program impact?
5. Given the questions above, what are the lessons learned regarding best practices in program evaluation involving student testing? How is this related to the development of project-specific or more general testing and evaluation systems?

The research team drafted these questions with all stakeholder groups in mind and with the understanding that information would come from several different sources.

## Methodology

This white paper is based on evidence gathered from a series of diagnoses prepared by the consultants to the CETTs and authors of this study, Dr. Gilbert Valverde and Dr. Richard Wolfe. Dr. Valverde and Dr. Wolfe provided technical assistance on assessment design and student testing to the CETT program from 2005 - 2009. It is important to note, however, that the consultants' involvement does not date back to CETT's inception. When implementation of the initiative began in 2002, a clear monitoring and evaluation plan was to be developed by each regional CETT; this was consistent with the goal of the initiative that each CETT would develop its own vision of program implementation and assessment. When technical assistance on student testing provided by the consultants started in the midst of Phase One, there were some elements of program evaluation already in place, though at various levels of development. Due to the lack of a well developed monitoring and evaluation plan from the program's inception, many early implementation decisions were made without consideration of the information and data needed to show progress and program impact over time. As a result, significant efforts had to be made by various evaluation teams in each CETT to compensate for this.

The authors reexamined their notes from former interviews and meetings, aides-mémoires, technical reports, and independent analyses to categorize the challenges faced in developing student performance tests in the three regional programs: C-CETT, Centro Andino, and CETT CA-RD. The authors were also extensively involved in assisting some of the countries with analysis of the test data. These analyses were reviewed and categorized to highlight examples that illustrate the fundamental, cross-cutting lessons learned in CETT, with the most important implications for future efforts.

## Limitations of the Research Study

The research team identified several limitations to this study:

- An innate limitation of this study is that information was taken from previous reports, discussions, interviews, and through the long-standing working relationships that the authors had with all three CETTs. As noted, the consultants who led this study provided technical assistance

on testing to CETT over the course of several years and were able to provide in-depth insight as to the inner workings of the testing initiative. At the same time, their close involvement with the program made it more difficult to analyze the program from an outsider's perspective. No new data were collected for this study.

- Though the research team provided technical assistance to all three CETTs, the level of involvement varied. The team was more involved with some of the CETTs, in particular Centro Andino and CETT CA-RD. Although efforts were made to review documentation, data, and data analysis across the three regions, the team's experience and access to documentation in some countries may have contributed to a more detailed analysis of those CETTs.

- Many topics in this white paper overlap other themes in the series. More in-depth analysis of these overlapping topics—regional nature, sustainability, and cost effectiveness in particular—is included in white papers one, three, and five respectively.

# Program Evaluation in CETT

The claim that an education program has impact – i.e., that it has a positive effect on the lives of students and that such impacts are superior to what would have been the case if the students had not participated in the program – is one that must be supported by clear standards of evidence. Evaluation is the systematic investigation of a program for the purposes of gathering evidence to determine its value.[3] A number of overviews of the assessment priorities of programs in Latin America and the Caribbean have stressed the importance of effective monitoring and evaluation systems to measure program impact.[4] Moreover, the development of indicators of program impact has been among the highest priorities of the strategic objectives in education of the LAC bureau at USAID, the implementers of the CETT program.[5]

In CETT, the monitoring and evaluation systems developed were intended to gather evidence of program impact through quantitative and qualitative indicators, including changes in student performance. This section outlines the design of the monitoring and evaluation system in CETT specifically related to student performance tests.

## Background

The aim of the Summit of the Americas Initiative, which was the conceptual basis for the CETT program, was to "improve teacher and school administrator quality and to improve the quality of reading instruction in the classroom throughout the hemisphere, with special emphasis on poorer countries and teachers who work in disadvantaged communities."[6] In addition to developing the five core components outlined in the Introduction, USAID also outlined specific impact measures for each of the CETT components. For example, three major results were identified for the teacher training component: [7]

- Teachers are more skilled, knowledgeable, motivated, self-confident, and better equipped to teach reading
- Fewer students are reading below grade level
- Networks of teachers and reading organizations are established and exchange best practices, lessons learned, and materials to improve reading instruction within their countries and across the hemisphere

With these and other impact measures in mind, each regional CETT was tasked with developing the CETT components based on the context of that region. In addition, according to the milestones set for the first year of implementation, the CETTs were to develop monitoring and evaluation plans within

[3] Rossi, P. H., Howard E. F., & Lipsey, M. W. (1999). *Evaluation: A Systematic Approach.* 6 ed. Thousand Oaks, CA: Sage Publications.

[4] World Bank. (1998). *Assessing Aid: A World Bank Policy Research Report.* New York, NY: Oxford University Press; Navarro, J. C., Taylor, C., Bernasconi, A., & Tyler, L. (Eds.). (2000). *Perspectivas sobre la reforma educativa: América Central en el contexto de políticas de educación en las Américas.* Washington, D.C.: U.S. Agency for International Development; Development Assistance Committee. (1999). *Criteria for Donor Agencies' Self-Assessment in Capacity Development.* Paris, France: Organisation for Economic Co-operation and Development.

[5] U.S. Agency for International Development, LAC Regional Office. (2002). *LAC Regional Education and Training Improvement Program Data Sheet.* Washington, D.C.: U.S. Agency for International Development.

[6] U.S. Department of State. (2003). *Centers of Excellence for Teacher Training in the Americas* [Press Release]. Retrieved from http://www.america.gov/st/washfile-english/2003/August/20030801114640nesnom0.254513.html

[7] U.S. Agency for International Development. (2002). *Centers for Excellence in Teaching Training: A Summit of the Americas Initiative Information Packet.* Retrieved from http://pdf.usaid.gov/pdf_docs/PNACY696.pdf

three months of signing the Cooperative Agreements with USAID.[8] In order to support this process, an external consulting firm, Aguirre International, was contracted to provide monitoring and evaluation support to the CETTs. A core part of the work of this consulting firm was to provide technical assistance in five areas:

1. Performance Assessment: Collaborate with USAID and the CETTs to develop a plan for monitoring key inputs, milestones, and program outcomes
2. Evaluation Research: Assist with indicator development, data collection and reporting, and trends analysis, drawing on the extensive work done in indicator development both by Aguirre International and other stakeholders for USAID in the arena of education in recent years
3. Planning: Work with CETTs to develop a methodology of "performance improvement" to develop and refine process, outcome and impact indicators, and to develop methods and strategies for appropriate data collection
4. Reporting and Dissemination: Assist the CETTs in creating the means to report efficiently the wide range of information of their activities
5. Training in M&E: Develop workshops and conferences to work with CETT partners to provide performance monitoring and evaluation training

It is evident that monitoring and evaluation was emphasized from the inception of the program, as resources were provided by USAID to support the CETTs in their development of performance measures. However, USAID and the CETTs soon found that developing valid and reliable measurement tools was a complex enterprise that required careful design and execution, and the complexity of evaluation would require a significant level of effort and time. Moreover, as discussed in other white papers in this series (see paper one: regional nature), the development of the CETT components themselves, such as the teacher training models, materials, and diagnostic tools, took a longer time than anticipated.

## Measurement Indicators and Demonstrating Program Impact

In the first two years of the program pressure increased to begin implementing the program and to demonstrate results quickly. The CETTs focused on measurement indicators related to the number of teachers and school administrators trained, and the relative number of students affected by the program. While these outcome measures were necessary, the CETTs were still developing measures of performance or impact. More focus was put on getting the program up and running before other pieces, such as performance evaluation, could be developed in a more holistic fashion.

It should also be noted that over the seven years in which the program was implemented, changes occurred in the assessment focus in terms of performance indicators. The first assessment tools created focused mainly on measuring *teacher* performance. In fact, two cross-regional qualitative studies of teacher professional development were carried out in Phase One (2004 and 2006). In subsequent years, the focus shifted to not only assessing teacher performance, but also creating instruments for measuring student performance in reading. This shift was in line with an increased focus by USAID—and the development industry—on measuring progress and the impact of program interventions on student learning outcomes.

Towards the end of Phase One the CETTs began to develop evaluation models that included student testing to measure impact at the program level. Common tests were developed by each of the three

[8] Ibid.

regional CETTs for use by the member countries in each region. Several years of hard work on the part of the CETT teams guided by technical advice from Dr. Valverde and Dr. Wolfe went into the production of the first comparative test results in 2007.

The C-CETT designed the Caribbean Reading Standards Achievement Test (CRSAT), a set of student achievement tests that measured growth in six key literacy areas. The CRSAT was based on the Caribbean Standards for Reading and Writing, which were developed and implemented by C-CETT. These standards were later endorsed by the Caribbean Community (CARICOM) and are currently used throughout primary schools of the English-speaking Caribbean.[9] The tests were administered to grade one to three students in C-CETT schools in Caribbean member countries.

In Centro Andino, grade three students in schools participating in CETT for three years were tested in all three countries at the beginning and end of the school year. In Peru only, the scores of these third graders were compared to the scores of third grade students in a chosen sample of comparison schools. CETT CA-RD also developed student achievement tests, and employed a pre-test and post-test design, in which CETT students of grades one to three were tested at the start and end of the school year. The structure and student outcomes of the student performance tests designed by each CETT are described in detail in the fifth white paper of the series on cost effectiveness.

It is important to highlight that the testing initiative examined in this paper is one part of CETT's monitoring and evaluation strategy. In addition to the external evaluation efforts carried out by Aguirre International (e.g., qualitative professional development studies, impact study, and other efforts), each CETT developed and/or received support to create their own internal monitoring and evaluation system. This system included, to varying degrees and forms in the different CETTs, tools to monitor the performance of teachers and trainers, diagnostic tools to map out students' levels at the start of the school year, and formative tests to evaluate students' progress during the school year. A number of these tools, where developed, became available in later years of program implementation.

[9] The Caribbean Community (CARICOM) is an association of 15 nations and dependencies throughout the Caribbean whose purpose is to promote economic integration and free trade among member states, and the coordination of labor, industrial, social, and foreign policies. CARICOM was established in 1973 by the Treaty of Chaguaramas (http://www.caricom.org/jsp/community/revised_treaty-text.pdf).

## A Fundamental Challenge: Evaluation Design

The plan for an evaluation effort is called a "design" and is the first step in providing an appropriate assessment of program impact. An effective design offers an opportunity to maximize the quality of the evaluation, helps minimize and justify the time and cost necessary to perform the work, and increases the strength of the key findings and recommendations by ensuring that threats to valid results are minimized. It represents a plan for the accumulation of evidence to substantiate all claims regarding the strengths and weaknesses of the educational intervention. Specifying the evaluation questions is of crucial importance, as is selecting appropriate methodological approaches.

In all regions, implementation of CETT began without a program-level evaluation design in place and thus no specific design for student testing. This proved to be a fundamental challenge throughout the program's existence. Consistent with the goal of encouraging each regional CETT to develop its own plan for implementation, the testing initiatives created by the CETTs varied in their approach. As a result, the regional assessment teams interpreted the broad imperative to assess "impact" according to their own understanding. This section notes three specific evaluation design challenges that affected CETT at the program level: 1) measuring program impact; 2) disaggregating the impact of program components; and 3) implementation of student testing.

### *Design Challenge 1: Measuring Program Impact*

CETT was primarily an initiative seeking to innovate in the area of teacher training. Much effort was put into developing an in-service teacher training model that would promote the use of effective teaching practices in literacy instruction with the assumption that improved teaching practices would result in enhanced opportunities to learn for children. Key assumptions of the model suggested a set of causal mechanics, a particular progression of changes in behavior that would lead to the final goal of improved student learning in reading and writing:

- Trainers help teachers acquire new pedagogical knowledge and skills through in-service training
- Teachers use new teaching proficiencies acquired in training to improve instruction, with the additional support of supervision and new instructional materials
- Improved instruction leads to better opportunities for students to learn
- Students take advantage of better opportunities to learn and acquire greater proficiencies in reading and writing than their peers in classrooms in which teachers did not receive CETT services

The final step in this progression, i.e., the link between opportunities to learn and greater proficiency among students, required a comparative judgment. Maturation in knowledge and proficiency in reading occurs, after all, in most of the world's schools, even those that underperform. What is more problematic is whether or not progression in learning in school matches the explicit curricular goals laid out for students. In order to substantiate a claim of program impact, the CETTs had to demonstrate that the students participating in the program had become more proficient in reading and writing as a result of the program intervention. The most important requirement was to demonstrate that the progress in average proficiencies in intervention schools was measurably superior to the type of learning that would occur if the CETT investments had not been made.

In order to accomplish this goal, the first requirement would be to identify the population of schools to which the CETT experience was intended to be generalizable and then assign schools to treatment or

comparison groups randomly. This is assuming that it would be possible to set up an experimental design for CETT testing. An experimental design was not part of the CETT program in the design phase because the testing initiative started after program implementation had already begun. Therefore, much effort was devoted to attempting to compensate for this fundamental challenge. The CETTs identified suggestive comparison schools after the program had begun rather than selecting ones at program startup. As a result there was no way to account for initial differences between schools, teachers or students, and inferences regarding program impact were difficult to substantiate. Much of the subsequent efforts to improve the testing design on the part of the regional teams, the authors of this report, and Aguirre International providing support to the CETTs in monitoring and evaluation were directed at compensating for this initial design weakness.

### *Design Challenge 2: The Impact of Program Components*

A second design challenge was the CETTs' ability to measure the relative impact of each program component, as several components made up the CETT package: teacher training, didactic and student materials, diagnostic assessment, etc. The hypothesis was that the introduction of a new practice would cause people to change their behaviors in desired directions and that these changes in behavior would cause desired changes in outcomes. However, the CETT schools received component interventions as soon as practicable, so teacher training, teacher and student materials, in-class coaching, formative assessments, etc., were introduced as soon as they became available. Monitoring and evaluation efforts provided crucial formative feedback to improve implementation, and efforts were made to refine the model, in particular in Phase Two. However, CETT did not test the different components of the program by experimenting with different mixes, and thus did not have a chance to refine the model by looking at the relative effectiveness of each component in comparison to the others.

### *Design Challenge 3: The Implementation of Testing*

A third challenge of the overall CETT design was that monitoring and evaluation activities were largely carried out by the implementing parties themselves. The efforts of Aguirre International helped compensate for this difficulty, in particular by providing expert advice and guidance in evaluation and testing design. Nevertheless, testing was in the hands of teams that had an important stake in the result of the evaluation efforts. Even when the best intention of all is to strive for objectivity, this presented problems of conflict of interest that could reasonably be invoked by external observers. In all cases, final decisions regarding evaluation plans related to testing and their implementation were in the hands of the regional assessment teams. Teams under the direction of CETT implementers were also the only responsible parties in data collection, analysis, and reporting. This is contrary to long-held standards of program evaluation.[10] Problems of potential or actual conflicts of interest should have been dealt with openly and honestly as recommended by the Joint Committee on Standards for Educational Evaluation.

As CETT evolved, considerations of data quality, professional standards in evaluation, and standards for reporting and communicating results became more salient. This expanding focus on monitoring and evaluation resulted in targeted support to the regional programs in their efforts to respond to these increasingly more explicit concerns. The next section provides some of the lessons learned and innovative techniques that the CETTs implemented in order to measure student performance given these design challenges.

[10] Joint Committee on Standards for Educational Evaluation. (1994). *The Program Evaluation Standards: How to Assess Evaluations of Educational Programs*. (2nd ed.). Thousand Oaks, CA: Sage Publications, Inc.

## Lessons Learned and Innovative Techniques in CETT Testing

In Phase Two of the program, Dr. Valverde and Dr. Wolfe drafted the Reference Standards document laying out in considerable detail all the steps and requirements for student testing, from test design to sampling, implementation, analysis, and reporting. The Reference Standards document was used by the CETT teams to address deficiencies in their test designs and to build knowledge about effective testing standards. The increased attention to internationally accepted standards in test design and implementation in CETT became a unique example of testing across national boundaries and a model for integration as an important aspect of results reporting.

Though complex and challenging, much can be learned from the CETT testing initiative. Considerable efforts were made to go beyond the types of perfunctory and superficial "evaluations" towards a more rigorous approach to and consideration of evidence regarding program impact using student performance tests. The most noteworthy characteristics of CETT evaluation efforts are noted.

1. **Use of pre- and post- measurement.** In order to measure program impact, Centro Andino and CETT CA-RD tested students in program and comparison schools over time. Since the schools were not chosen randomly into intervention and control schools prior to the program implementation, the CETTs chose comparison schools with similar characteristics as the CETT cohorts. An advantage of having the same measurements at two points in time was that the CETT testing teams could then measure the amount of change in the outcome or performance variables, what is termed in evaluation terminology as the "value-added". These techniques were innovative and should be recognized as important contributions to program evaluation.

2. **Vertical scaling across grades**. Vertical scaling, described later in this section, further strengthened the choice of pre-post measurement design in Centro Andino and CETT CA-RD. Growth and learning are fundamental assumptions of education. The premise is that school children progress in their learning over time and across the grades. Despite the centrality of this philosophy and its acceptance by most educational system actors, few efforts in program evaluation incorporate this perspective in the evaluation design. CETT took up the challenge of mapping and locating students along a learning continuum from grade one to grade three.

3. **Multi-year tracking of teachers and students, to look at incremental effects of training and instruction.** As the evaluation efforts in CETT were refined, extended, and strengthened, efforts were made to further enhance the learning progressions perspectives referred to above. Thus, some of the same students and teachers were tracked and studied over time in order to gauge the progressive impact of the implementation of the CETT model. This resulted in a true longitudinal study, not simply a succession of independently sampled groups over time.

4. **Test developments that were aimed at studying reading achievement through the lens of the program objectives.** Although tests in CETT, with the possible exception of C-CETT, were essentially norm-referenced,[11] there were noteworthy efforts to design measures aligned with the learning goals that served as the objectives of the program. Alignment was sometimes problematic, both because consensus on the program's pedagogical models was achieved over time and because it was evolving simultaneously with early efforts in test design. However, analysis and reporting emphasized use of the program objectives.

[11] In everyday language, "norm-referenced" testing is often thought of as "grading against a curve."

5. **Interpretation of item and sub-domain results for diagnosis**. CETT took on a more nuanced and rigorous analysis of test data than is often the case. Rather than being satisfied with general measures of achievement in reading comprehension and early reading skills, CETT analyzed student performance on individual items or sub-scales of items measuring discrete skills or pieces of knowledge that made up the larger domain of reading comprehension. Thus, it was possible for CETT, especially in the final analyses conducted, to not only gauge overall levels of proficiency, but to also identify areas of weakness and strength for diagnostic purposes.

6. **Collection and analysis of associated variables for students, teachers, and schools.** Evaluations in CETT attempted to account for the effects of demographic, social, and other contextual factors on what students learned. Analytical models in CETT evaluation were therefore capable of accounting for these associated variables, in order to better isolate CETT effects from the impact of other external factors affecting outcomes.

Tests, surveys, observation protocols, and other evaluation instruments should embody the operational definitions of program goals. Data from any evaluation are only as good as the questions asked, achievement items posed, and other verification strategies followed. Poor instruments result in data that cannot be used to sufficiently substantiate claims regarding the impact of an educational innovation. The following two sections identify innovative techniques that were used by the CETTs in developing their student tests and in analyzing test results.

## Innovative Techniques in Instrumentation

The area that received perhaps the most attention in CETT was instrument development, or the development of student achievement tests. It was the area in which the largest number of CETT personnel participated and that required the greatest efforts of coordination across units. Initial designs were refined over the years, and there was an increasingly rigorous use of pilot data. External feedback and coaching on test development was limited, as were efforts made to set cut scores or achievement levels as outcome goals. However, as efforts progressed, the CETTs developed increasingly stronger technical understandings of the attributes of good test items and some procedures for assessing and validating the quality of the items.

### *Test Development and Matrix Sampling*

The construction and interpretation of student achievement tests used in the context of program evaluation is often oversimplified. If the questions on a test are *about* the content of the target instruction and learning, then one may make the conclusion that one test will be as good as another. However, this kind of thinking has negative consequences for making inferences about program quality:

1. If a test does not provide *comprehensive* measurement of the target content, then it will be difficult to determine the overall effect of the program. For example, if a test used in the evaluation of a reading program measures only decoding skills, one will have little information about the impact of the program on student reading comprehension and inference.
2. If a test does not provide *differentiated* information about student achievement across the target content, then there will be no specific diagnostic information for improving the skills of individual students and no formative indications for improving the instruction and the program.
3. If a test does not provide information that is *referenced* to criterion standards of achievements, then it will be difficult to judge whether a program has succeeded. For example, in the analysis of reading comprehension, one should want to know whether students can adequately

understand and use the texts found in specific populations of reading materials, such as schoolbooks.

First, one must recognize that achievement content domains are large and complex. In the case of reading in early grades, the domain for testing can be analyzed in various ways. One may ask what skills students have (decoding, vocabulary, finding explicit information, making inferences, making connections); try to measure the reading experiences of students and the interests they show; and inquire about what students can do with the reading activity, such as discuss, write, and communicate the main ideas.

However, all these considerations lead to a crucial methodological problem: if the content domain is large and complex, then the tests used to evaluate student achievement in the domain must be comparably long and detailed. A test with complete, detailed coverage of a content domain would be much too long for one student to take in a reasonable testing session. One way to overcome this problem is to use matrix sampling. In matrix sampling, all of the test items are randomly divided to create different tests. Then alternate tests are distributed randomly. Having multiple forms also makes it practical to administer effectively the same test at two points in time, such as at pre-test and post-test, since the situation of having students see the same items twice can be minimized or discounted. The text box below includes more detailed information on matrix sampling.

**Technical Example: Matrix Sampling**

A methodologically sound way to overcome the problem of too much information to be included in one student exam is matrix sampling, where the pool of items necessary to provide full coverage of the content domain is divided into a number of alternate test forms, each containing a stratified random sample of the pool. At each critical point in a study (e.g., pre-test, post-test), the alternate forms are distributed randomly, one to each student across classes and schools. The statistical situation is then:

(a) Each student has taken a sample of questions that is representative of the total pool—that is, of the content domain—so the student's score estimates the student's performance in the domain. The scores may not be very precise, because an individual's sample is small, but they are unbiased, so correlations with external variables will be correct (after correction for the sampling).

(b) The aggregate scores over the students in a classroom, or a school, or a treatment condition (e.g., experimental and control) also estimate performance in the total domain. They are quite accurate because they are determined by averaging over all students and therefore over all forms and all items, and the sampling errors in the test forms of individual students cancel out.

(c) At the level of individual student, there is likely insufficient information (too few items) to form analyzable scores for sub-domains. But between students, sub-domain scores are determined by different items due to forms-level sampling. Consequently, once aggregated over the students in a class, or school, or treatment condition, sub-domain scores can be quite accurate.

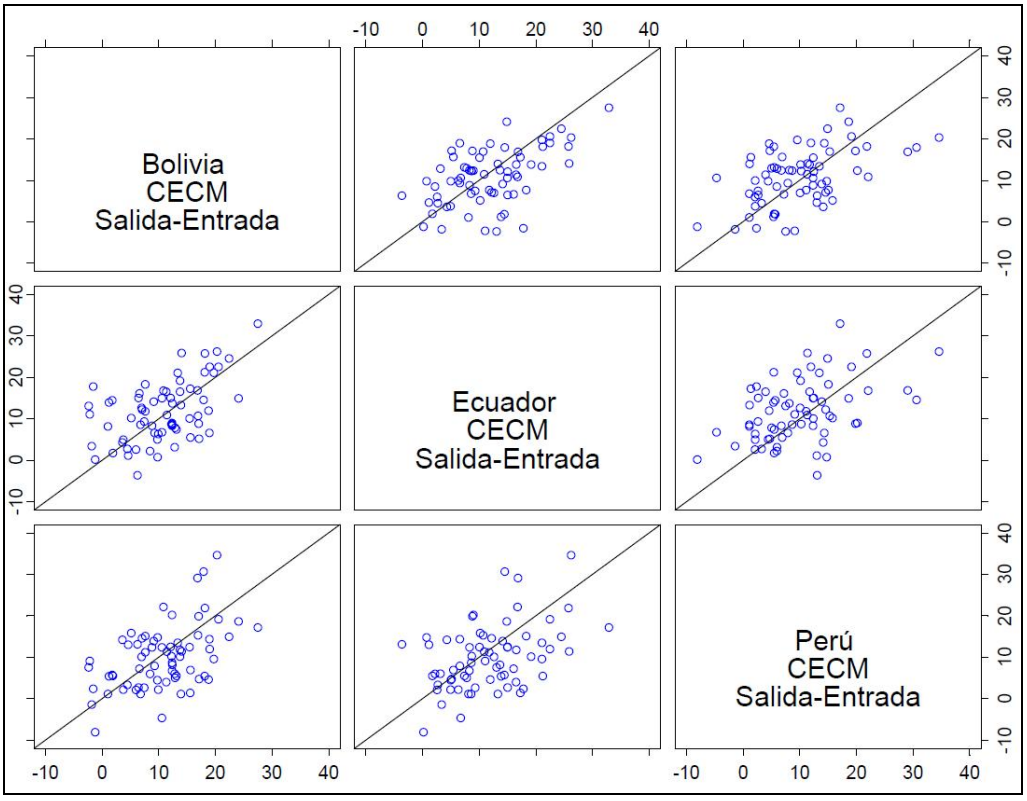
In Centro Andino, matrix sampling was used for testing in the 2008-2009 data collection round. As shown in the following table, the test design comprised four booklets, each divided into two blocks, and each student answered one booklet.

| **Matrix Sampling in Centro Andino: 2008-2009** | | |
|---|---|---|
| Booklet | Block A | Block B |
| 1 | C1a | C1b |
| 2 | C2a | C2b |
| 3 | C3a | C3b |
| 4 | C4a (*) | C4b |
| * This test was from 2006-2007 | | |

Block A in the booklets contained most of the reading comprehension items and Block B contained most of the text production items. Note that Block A in Booklet 4 was copied exactly from an earlier test. Because all booklets were taken by a random sample of the students participating in testing in 2008-2009, comparisons could be made with the earlier results. The new items in the other seven blocks represent samples across the content domain. The number of items in the total testing pool is 3.5 times larger than what would have been available if each student took exactly the same test. This allowed much more detailed and accurate comparisons of achievement, especially between beginning and end of year, participating countries, and treatment and comparison schools.

Display 1, taken from a Centro Andino report, illustrates the benefits of using matrix sampling. Each graph compares a pair of countries and each point plotted is one item. The location is determined by the pre-test (*entrada*) to post-test (*salida*) gain in student achievement with the vertical axis corresponding to one country and the horizontal to another. These are data for grade three in Centro Andino schools (called Centro de Excelencia para la Capacitación de Maestros, or CECM, in Display I).

**Display 1: Inter-country comparisons in item difficulty changes, 2009 Centro Andino**



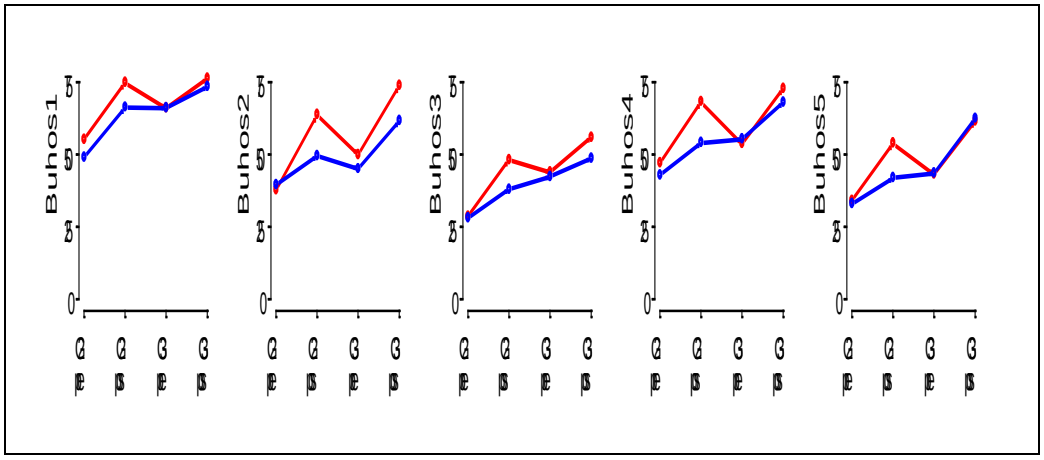

The gain is nearly always positive. For many items, the growth is larger in Ecuador than in Peru; in the bottom center graph, most of the points are below the diagonal, which marks equality of gain for the

two countries. Furthermore, there is a substantial correlation of gains: items that show more gain in Ecuador show more gain in Peru. At the same time, there is a lot of scatter, corresponding to variation over items in the relative amount of gain in Ecuador and Peru. Tracking this down may reveal differences in program implementation, specific language and cultural differences, and/or differential emphasis and accomplishment in the aspects of the reading program. The advantage obtained from matrix sampling is that there is a large number of items available in the pool to allow for finding trends and tracking interactions and to provide highly accurate summary comparisons.

### *Vertical Scaling*

Other refinements to test design took place as the three CETTs expanded their testing efforts. In CETT CA-RD, for example, the assessment team designed a test to measure vertical scaling on reading comprehension items between grade two and grade three. As seen in Display 2 below, the test given to CETT (red line) and comparison (blue line) schools consisted of a reading passage (called "Buhos") and five corresponding questions. Since the samples of students were random, the difficulties of the items could be compared across the beginning and end of the school year (shown as "pre" and "pos"). The graphs indicate how the difficulty of each item changed from the pre-test to the post-test in the Dominican Republic. The interest for interpretation lies in the variations in growth patterns. These can be linked to the specific contents of the items.

**Display 2: Vertical scaling of test items in grades two and three: Dominican Republic**




In most cases, the within-year gains in CETT schools were higher (the slopes were greater) than in the comparison schools. While the starting points for CETT and comparison schools were about the same in grade two, the startling result is that there seemed to be a drop in achievement in CETT schools between the end of grade two and the beginning of grade three. This was a similar finding in the CETT impact study (2008-2009), which found that competencies among CETT students decreased during the summer months.[12] However, this dip could also signify a design flaw; as this was a cross-sectional design, the student samples were chosen from two separate grades in the same school year.[13]

[12] Aguirre Division of JBS International, Inc. (2009). *Centers for Excellence in Teacher Training (CETT): Two-year Impact Study Report (2008 – 2009).* Retrieved from: http://pdf.usaid.gov/pdf_docs/PDACS248.pdf

[13] Cross-sectional data refers to data collected on subjects (in this case data on students) at the same point in time, or without regard to differences in time; it differs from longitudinal data, which follows the *same* subjects over time.

### *Measuring Test Validity*

Assessments are conducted with the aim to discover, describe, and interpret facets of the educational program they assess. The strategy followed in CETT involved testing students to demonstrate that they had acquired a number of skills expected of them. The assessment system involved a *written test* with questions that, in the view of the test's authors, demanded that the students use what they learned from CETT-trained teachers in order to answer correctly.

Interpreting the test information correctly and using it properly demands an understanding of the kind of representation of achievement or learning that the tests allow. That in turn means attending to what is known in educational measurement as *validity*. Validity is not intrinsic to the tests, but instead is a property of the interpretations of the information obtained through them and the uses to which that information is put. Hence validity is currently defined as the degree to which empirical evidence support the interpretation of the results of an assessment.[14]

In the case of achievement tests, whether they are norm-referenced, as was the case in CETT, or criterion-referenced, efforts are made to draw conclusions that go beyond the test questions.[15] In other words, in both cases it is acknowledged that the test questions account for only a small sample of all the possible questions that could be asked in an effort to determine if the students have acquired certain abilities. The conclusion drawn from analysis of the kinds of tests mentioned is that if the students give correct answers to 80 percent of the test questions they could correctly answer 80 percent of all the questions that could possibly be asked to assess that ability.

Of the utmost importance in validating test results is that the technical designs of the assessment clearly address the following:

- **Abilities and skills.** In CETT, a variety of efforts was made to specify the abilities and skills to be assessed. InC-CETT, a set of common standards was agreed upon as the definition of the learning goals of CETT. These Caribbean Standards for Reading and Writing were used as the referents for the student tests.[16] A contrasting model was followed initially in Centro Andino. In this case, external consultants were asked to put together the first tests, and these were based primarily on test item expertise that these consultants had developed in working on the Peruvian national tests. This resulted in fairly weak connections to the specific programmatic goals of CETT. In CETT CA-RD, a number of test specifications were drawn up with different lists of "competencies" or domains to be assessed, which were vetted by the teams of the participating countries.

- **Consistency between the questions and the abilities or skills to be measured.** This was a weak area across the CETTs, especially in their early efforts. At first, in all cases none of the procedures to ensure consistency were specified, and only as efforts progressed, were procedures for rigorous review attempted. However, these efforts at review, in the case of CETT CA-RD for

[14] For a long time, the most common and extensive concept of validity, and one that dominated academic thinking and assessment practices, was that proposed in 1949 by L. J. Cronbach in his book *Essentials of Psychological Testing* (New York, NY: Harper and Row). Since then, the evolution of the theory and methods of psychological and educational assessments has given rise to a new conceptualization and to its standardization among professionals in these fields. The third edition of *Educational Measurement* by R.L. Linn (1989) presented Samuel Messick's proposal that established the current thinking. Revisions of this proposal led to the meaning of validity as documented in the Standards for Psychological and Educational Measurement.

[15] As noted earlier, "norm-referenced" testing is often thought of as "grading against a curve." "Criterion-referenced testing" refers to an individual's ability to answer questions posed correctly, e.g., as on a driver's license exam, regardless of how well or how poorly other people being tested perform.

[16] As noted earlier, the standards developed by C-CETT are currently implemented throughout primary schools of the English-speaking Caribbean.

example, took place simultaneously with instrument pilots, so changes, corrections or refinements resulting from these were not piloted, substantially constraining their usefulness.

- **Types of questions that demonstrate the abilities mastered.** Discussion of these questions, and developing strong technical answers, were not initially part of the CETT evaluation efforts. As CETT progressed, however, there were efforts to experiment with different types of measures. For example, CETT CA-RD developed items that attempted to measure pre-reading skills, and tested open-ended items to attempt to measure writing/composition skills. In neither case was there enough time and effort to validate these measures, given that this came later in the process. In the case of the open-ended questions, a viable analysis plan was never developed and the data were never used. Tests in Centro Andino paid little attention to improvement in development and design of items and definitions of ability indicators. Technical documentation on C-CETT procedures in regard to validity were not made available to the testing consultant team.

- **Students' chance of demonstrating what they know is not affected by factors beyond their control.** It is important to describe how it was ensured that all students have the same chance to demonstrate what they know. This aspect of validation showed the most substantial improvement in CETT CA-RD. Using Item Response Theory (IRT) scaling methods, pilot data were used to identify items that showed substantial item-by-country interactions that suggested significant if unintentional bias. Such items were eliminated, resulting in a more valid set of measures.

It must be understood that validity is a question of *degree*. No measurements are perfectly valid, and none faithfully reflects every aspect of the educational circumstances that they seek to measure. Some measurements, however, are more or less valid, depending on the conclusions that are to be drawn from them or the use to be made of the information they provide. For example, factors such as living in rural versus urban areas can (but do not necessarily) result in significant differences in test results. In addition of being aware of such factors, educators should try to ensure that curricula and test materials also do not unduly disadvantage children from one locale over those from another.

The responsibility for validating assessments is borne by both designers and users. Those who design assessments have a responsibility to explain clearly what they are and are not useful for, and should report all relevant information so that users have grounds for judging the validity of the assessments. Users have a responsibility to use the results in line with their validity criteria; if users propose a new use for the assessments, they have to validate them for that new purpose. Since the goal is to ensure congruence between the assessment and the educational circumstances being assessed, validation is a *scientific* activity. It is also a *technical development activity* because the task of gathering evidence of validity often spurs the redesign or fine-tuning of the instruments or their theoretical bases. Toward the conclusion of CETT, especially in CETT CA-RD, efforts to use validation activities to improve the assessments were increasingly included as part of the evaluation process.

## Innovative Techniques in Test Analysis

Analysis is disciplined inquiry into the links between indicators of program elements and educational outcomes. Data by themselves do not prove or disprove claims about program impact. Analysis converts data into evidence substantiating claims of this type. As noted, the testing initiative came into play at later stages of program implementation in CETT, and the development of student achievement tests to be used across participating countries was a complex and lengthy process. These realities explain why analysis plans were not prepared in advance. This constrained efforts to optimize

resources, such as time, technical expertise, and statistical software in carrying out analyses. This section presents analysis issues and lessons learned.

### *The Importance of Scaling and Equating*

In CETT, the assessment designs hinged on comparisons over time, especially looking at the growth achieved from one grade to the next and from the beginning of one school year to its end. In some participating countries, the same test was used repeatedly. This can be problematic because repeated testing can cloud interpretation of results: For example, is improvement in scores due to recollection by students of texts and questions they have already seen? Do teachers use testing materials as part of instruction in ways that would cause scores to rise without valid generalization to other texts and questions?

The correct technical response to these problems is to use "parallel" but not identical tests in repeated measurements. One way to make tests parallel was mentioned earlier: Two or more tests can be constructed as stratified random samples from a single matrix of texts and items, and the scores will be automatically *equated*, at least up to the limits of (item) sampling error. As noted, this strategy increased the total number of items and overall accuracy of class, school and program measurement, and allowed equivalent scores to be obtained for individuals. The individual scores may have a good deal of sampling (measurement) error, but will be unbiased and useful for relational analysis.

Randomly stratified parallel forms also can be used over time, with a different form or set of forms used at each time point. The scores from different times will be comparable, both at the level of individuals and, with much greater accuracy, for classes, schools, and programs. In order to accomplish this more effectively, the following should be considered: (a) a large pool of items are needed to be randomized into as many forms as are needed, within one time point and within and across grades; and (b) it must be recognized that the scores from different test forms are not exactly equated but rather differ statistically because of item sampling, and further that the samples are quite small (maybe as few as 20 questions per form).

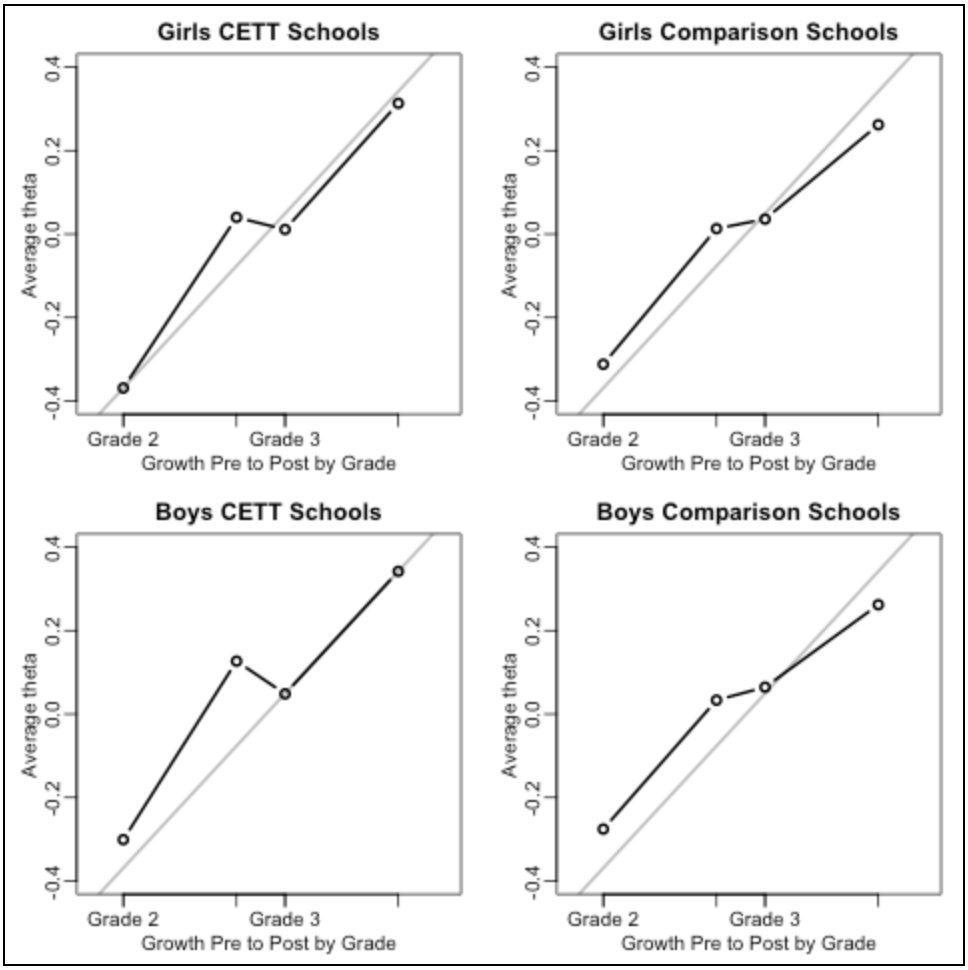
An efficient measurement solution is to use Item Response Theory (IRT) programming.[17] When IRT analysis is successfully applied, it provides a comprehensive measurement system for dealing with multiple test forms over time. In addition to having stratified random selections of items over the different forms, one includes blocks of items that are common to forms. For example, in the last CETT CA-RD test design for grades two and three, about 40 percent of the items were common to the different forms and 60 percent were unique to forms. This meant that the total item pool was reduced somewhat but the advantage was obtained that all forms from grades two and three could be scored on a single scale. The common items were used for the IRT equating, and all items contributed information to individual, class, school, and program scores.

Display 3 is an example of the results from CETT CA-RD. The vertical scales are in the IRT metric and labeled "theta." This represents the estimated latent ability or achievement in a way that is comparable across the two grades. The graphs plot the aggregated growth trajectories for students, divided by sex and program type. These are now true longitudinal findings, unlike the earlier graph that was

[17] IRT is a *theory*, one that postulates a single latent trait or ability that differentiates students. Student responses to items are hypothesized to reflect their latent abilities according to particular functions that derive from item characteristics or *parameters* of difficulty and discrimination. Limitations to be noted: If items and patterns of test response do not follow the theory—e.g., if there are multiple factors or if there is a lot of random guessing—then the theory and its calculations can be invalid. Second, the computations for IRT analysis are not simple. They require specialized software and sometimes difficult and sensitive computing steps.

constructed from cross-sectional data. The grey diagonal simply marks the potential linear growth from the lowest to the highest mean. It is clear that for both boys and girls, the CETT schools started at a lower score and ended up higher than the comparison schools. Within each grade, the growth for CETT schools was greater. As also noted earlier, the graphs showed that the CETT schools dropped in achievement results between the end of grade two and the beginning of grade three.

**Display 3. Longitudinal growth in overall reading from 2008-2009: CETT CA-RD**




***Providing Detailed Information About Achievement***

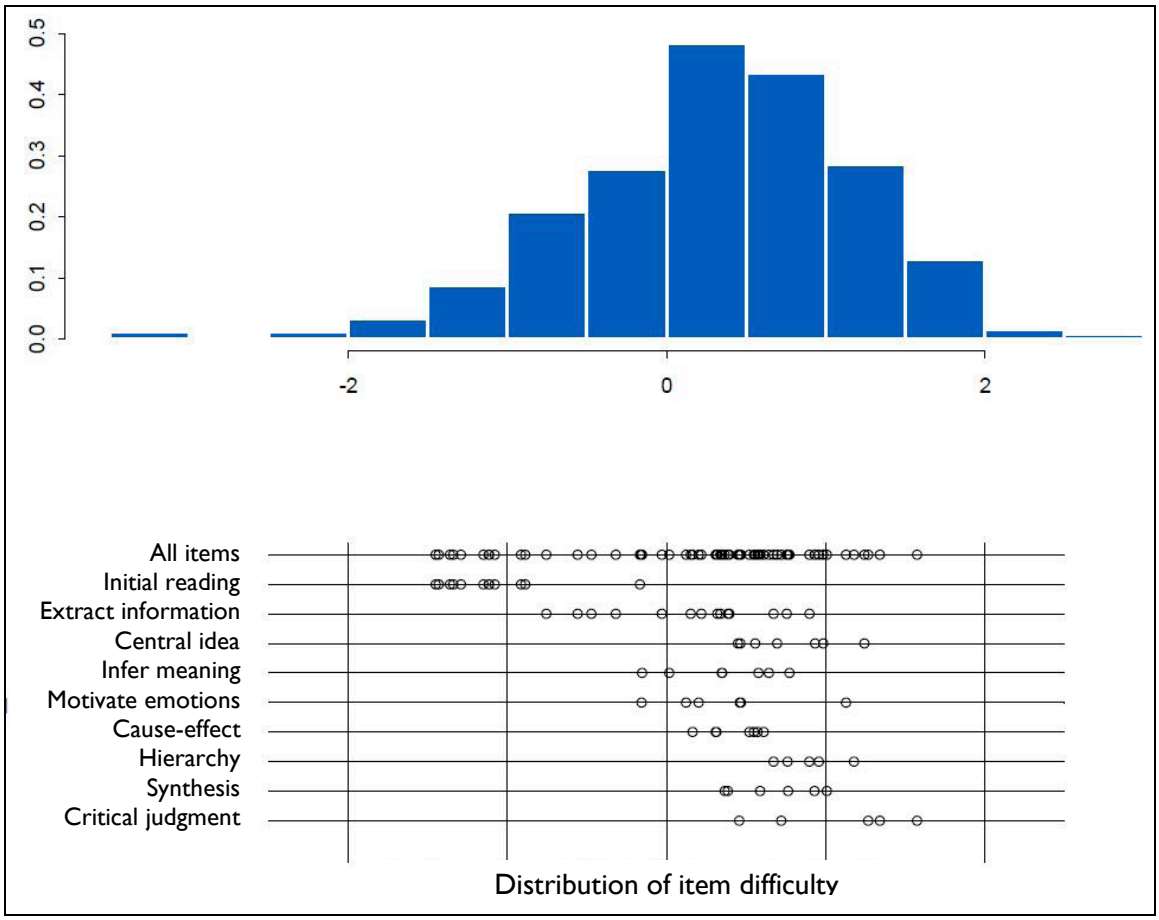
It is argued above that one should use a test design that provides enough items so that many parts of the content domain are represented in the measurement. An additional and important goal is to provide *differentiated* information about sub-constructs within the total content domain, and to what degrees these sub-constructs are learned by different students, in different classrooms. A typical division of reading comprehension is into the sub-constructs of finding explicit information, making inferences, and making connections. For illustration, the items given as points in Display 1 and lines in Display 2 were identified with their sub-constructs to see how much of the variation is explained by content and ultimately by population and program differences. These analyses are given in Display 3, which attempt to determine if the trends are diminished or amplified in different content areas.

At the program and group (school, class) level, this is in theory possible, at least for those sub-constructs that have sufficient numbers of items overall represented in the test, especially since it is possible to pool information over all forms. Around 15-20 items are needed to adequately measure a construct, and that is true for each construct examined here.

At the level of individual students, the situation is very different, since with a total of about 20-25 items per student, no sub-construct is going to have a sufficient number of items to justify a sub-score at the individual level. This is not just a matter of the unreliability of estimating an individual's sub-score, but more importantly, the sub-scores would not be consistent between individuals with different forms.

One approach to understanding the content differentiation within a test is based on the IRT analysis, presented in the form of a "Wright Map," as in Display 4. The top part of the display shows the distribution of the overall achievement score, given as an IRT theta. The bottom part of the display shows the location of each item in the test on the theta scale, where that is determined as the theta score necessary to make a correct response likely (80 percent probability). It is immediately interpretable that students with thetas of less than 0 can do only the items of initial reading with any certainty; about half of the items concerning extracting information are accessible; and virtually none of the other contents. On the other hand, students with theta score of one or more can answer most of the items correctly, including half of the most difficult items concerning critical judgment.

**Display 4: Wright Map for the 2009 Centro Andino: Test in Grade 3**



Not very much attention was given in any of the CETT analyses to sub-score results. This was probably not due to a lack of interest but rather to complexities in the measurement and the analysis methodology required.

### *Other Opportunities for Evaluating Achievement Results*

The root meaning of "evaluation" is to assign or infer *value*. This is not a trivial issue, because achievement tests, especially when constructed from multiple-choice questions, do not automatically or easily lend themselves to judgments of adequacy, or sufficiency, or goodness of achievement. In fact, it is easy to construct in the same content area, or with the same item pool, an easy test or a hard test, simply by picking items with low or high item difficulty parameters. The *percent of correct responses* on a test is commonly thought to be a meaningful indicator of accomplishment, with some accepted cut-score (e.g., 50 percent, 70 percent, 90 percent). But the usefulness of that is easily seen to be spurious; it all depends on the test and the items.

The issue being discussed is judging when student test scores demonstrate attainment of an *educational standard*. The idea is to convert test scores, which have an essentially arbitrary metric, whether it is percent correct on a test form or an IRT theta score, into the performance levels that can be identified with attainment of an educational standard. For the purpose of program evaluation, the intention is to be able to say that a certain percentage of the students reached an adequate level of performance on the desired standard for reading comprehension, perhaps going further to identify what percentage reached a proficient or advanced level of performance.

There are a number of methodologies for collecting and reconciling judgments about test performance to convert test scores into performance levels corresponding to an education standard.[18] This is a complex and controversial area of research and practice in educational studies. In the case of CETT, more time and effort would be needed to be able to express results in terms of standards. Again, unfortunately this strategy was not built in the evaluation design from the start, perhaps because the standards for student achievement were not ready or not agreed upon then, or the methodology for standard setting was not invoked. More work on this aspect would be valuable, since having only raw or relative results makes it difficult to decide how much success was ultimately obtained.

[18] Cizek, G. J. (2001). *Setting Performance Standards*. Mahwah, NJ: Lawrence Erlbaum Associates.

## Recommendations

Based on the CETT experience, this section puts forth four overall recommendations for designing program evaluation and student testing efforts in future educational programs. The recommendations are intended to advance discussion regarding directions, strategies, and research practices in similar initiatives wishing to measure program impact using student performance tests. It is recognized that accomplishing all of these goals may be a challenge depending on the country context. These recommendations are included as best practices from a research standpoint, and should be taken into account as much as local conditions allow.

***Recommendation 1: Design evaluation components, such as student testing, in advance of program implementation.***

The major conclusion of this paper is that the evaluation design of programs such as CETT must be carefully planned and executed with diligence, and ideally should be in place prior to implementation startup. It is recommended that the design include: 1) a complete specification of the evaluation questions under study; 2) identification and justification of methodological strategies for answering these questions; 3) a data collection plan that anticipates and addresses problems that may be encountered; 4) an analysis plan that will ensure that questions are answered appropriately; and 5) a description of the anticipated reports. Specifying the evaluation questions and selecting appropriate methodological approaches is crucial.

The overall design of a testing initiative should be planned in advance for a realistic sense of the size of the intervention (treatment) effects that are likely to be obtained, recognizing that most treatments provide modest but important improvements. Therefore, sample sizes should be carefully considered, especially numbers of teachers and schools, so that the samples will provide sufficient statistical power for demonstrating effects of those magnitudes. Additionally, sufficient time for teachers and schools to learn and adjust to new methods and for students to gain knowledge and skills should be allowed. This usually involves more than one school year. The design should also have the expressed purpose of fully taking advantage of all sources of variation. For example, if more than one country will serve as an implementation site for the initiative, cross-national comparisons should be made.

***Recommendation 2: Design a scientifically sound plan for how and where the intervention and its comparison are to be implemented, taking into account from the beginning the requirements to ascertain potential program impact.***

A goal of future work should be to rigorously consider educational initiatives as experiments and efforts to test causal hypotheses. Therefore, the selection of intervention sites and comparison or control sites to take part in testing should be done with the primary goal of obtaining generalizable results from the experiment. The first step should be to identify the population for which results are intended to be generalizable, and then randomly assign intervention and control units. These samples should be appropriately stratified to increase generalizability of results. Ideally, the implementation will also involve replication over time.

All elements of the intervention should be evaluated; therefore, different mixes of elements should be tested in different randomly designated intervention sites in sufficient numbers. Only in this way will a testing team be in a position to determine the relative value of, for example, teacher training and student assessment versus new student textbooks. Each of these elements represents the use of important resources, and overall, evaluation and monitoring systems should be in a position to

determine their relative contributions to desired outcomes, with an eye to optimizing investments in the future.

***Recommendation 3: Specify the design of the measurement of achievement in the evaluation plans. Include the specification of the procedures that will be followed to ensure that such testing measures are valid and reliable and can be used comparably across time and conditions.***

This aspect of evaluation designs should include specification of the procedures that will be followed to include comprehensive coverage of the content domains that are relevant to the instructional goals of the project or program. They should include provisions for vertical alignment of measurements across grades when interventions are for multiple grades. They should also include the plans for the construction and validation of measurement scales that allow repeated measures.

***Recommendation 4: Go beyond investigating causal effects. Design ways of investigating and confirming the quality of the implementation of the project objectives, and assess their relative impact on outcomes.***

Evaluation designs should include laying out procedures for assuring that the content of materials and instructional training are measured in correspondence with program objectives. They should include measures of teacher learning and practices if the program focuses on helping teachers learn new ways to teach. They should also envision measures of opportunities to learn, such as instructional practices, and make provisions to link all of the "process" measures to the measures of student outcomes. These measures should be compared to the control or comparison sites, in order to rigorously evaluate program impact.