Standard Setting Issues and Practice in Latin America

Michael Rodriguez, University of Minnesota

Fernando Rubio, Juarez Associates, USAID–Guatemala

Jeff Landsdale & Zarko Vukmirovic, American Institutes for Research, USAID–Honduras

Lorena Meckes & Jacqueline Gysling, Universidad Católica, Chile

With assistance from

Luis Ramos, Ada Luz Pineda, Mariano Solorzano, Mauricio Estrada, & Maura Rivera, Honduras

Otto Rego, Guatemala

April 10, 2011

(Version Date: April 8, 2011)

Paper presented at the symposium on Standard Setting in an International Context: Issues and Practice, at the annual meeting of the NCME, New Orleans, LA.

Standard setting issues and practice in Latin America

Much of the language, ideas, and challenges of large-scale assessment are new to the countries of Central America and relatively new in much of South America. The challenges are particularly acute in the process of standard setting, involving key stakeholders who have only recently been introduced to the idea of national academic standards, national assessments, and the idea of setting performance standards. This presentation summarizes the thinking behind the planning, the experiences, and the lessons learned through the analysis of standard setting procedures, results, and feedback in three contexts, including Chile with the most experience and ongoing challenges of reporting results in meaningful ways; Honduras with standards based assessments since 2005 and their ongoing challenges of exploring school factors that explain variation in outcomes; and Guatemala with the most recent standard setting experience, first introduced in 2008, facing intense challenges regarding stark disparities in opportunities to learn across the country. We will also focus on the use of language and introduction of the ideas of standard setting as they impact the work of the standard setting panels across the three contexts and over time.

***Guatemala**: Fernando Rubio, Juarez Associates*

Through early USAID efforts in strengthening basic education in Guatemala, The Universidad del Valle de Guatemala and the Ministry of Education created the National System for Measuring Academic Achievement, which was renamed the National Program for School Achievement Assessment (PRONERE) in 1997, with additional support from the World Bank. Until 2001, national assessment activities were carried out by the University. USAID funding in 2005-2009 (and continuation funding) supported the development of new academic content standards and associated national assessments in reading and mathematics. In some grade levels, the assessments are samples, in others they are a census. The tests are designed to be aligned with the newly developed content standards and to facilitate longitudinal analyses and growth modeling. The USAID-Guatemala project has worked to develop the current testing system through IRT, to facilitate strong scaling and equating. These developments have provided opportunities for training of staff within the Ministry on modern measurement theory and advanced statistics. The assessment system still faces challenges, including the multi-language assessments in early primary levels, limited opportunities to learn, and limited teacher preparation and professional development.

Honduras*: Jeff Landsdale & Zarko Vukmirovic, American Institutes of Research*

National assessment began in Honduras during the early 1990s, through education improvement projects with USAID and the World Bank. The recent and current USAID-funded projects are an integrated set of standards and testing development and implementation, technical support, and capacity building activities that address major components of the Honduran Ministry of Education's national education program. AIR collaborates with local educators in assisting the government to develop content standards, teacher guides, teacher training modules, strategies for timely use of test data, supervisor support systems, and a standardized testing system to increase student learning outcomes for meeting Education for All goals on student achievement. Honduras is now in the process of setting up its own para-governmental National Assessment Institute, to ensure the long-term sustainability of systems built under the project. Significant third-party financial support has been leveraged from private and public sources to make sure that appropriate levels of funding is available, thereby also stimulating a wider involvement of the broad stakeholder community to support a standards-aligned educational reform movement.

Chile*: Lorena Meckes & Jacqueline Gysling, Universidad Católica*

National assessments have taken place in Chile since the early 1980s. The current national assessment program is administered by SIMCE (National System for Assessment of Educational Quality), including experimental samples at some levels and national census at others. The existing assessment system was developed and initially administered by the Catholic University of Chile, but has since been integrated into the Ministry of Education. Assessment results have moved from informing the public to developing educational policy. Chile is currently undertaking establishment of standards and refining assessments. The summary of standard setting in Chile contains excerpts from a more comprehensive report, which is available from the authors.

Standard Setting in Guatemala

The United States Agency for International Development (USAID) executes its Guatemala projects inside the framework of its Regional Strategy for Central America and Mexico. USAID-Guatemala reinforces local efforts in the educational arena, supporting the government of Guatemala, through the Ministry of Education (Ministerio de Educación, MINEDUC) and civil and social organizations. The goals of these efforts include improving the transparency, efficiency, and effectiveness of the educational system; achieving universal access to primary education; and increasing educational quality.

To support the improvement of the efficiency, equity and quality of the educational system, USAID supported, through a 4 year (2005-2009) grant, the Educational Standards and

Research Program. This program, administered by the firm Juárez and Associates, provided technical and financial support to the MINEDUC, utilizing results of educational research and evaluation activities. In addition, the program developed an active communication and dissemination process that informs the national dialogue on education. An important aspect of this project was the development of the National System of Assessment and Evaluation.

A first step in this project was the creation of a national set of content standards in several areas of the curriculum for K-12 general education public schools. To begin the process of monitoring performance on these new content standards, national standardized assessments were developed in the areas of Mathematics and Language Arts. Following the first operational administration, a low-stakes administration without consequences to students or schools, provisional performance standards were set. Following the first complete administration and analysis of performance levels, additional changes were made to the content standards, requiring subsequent changes to the tests and the need for new performance standards to be set. All of this was by design, as Guatemala has never had national assessments for accountability purposes.

Guatemala has a population of approximately 14 million people and more than 50% speak one of over 20 Mayan languages as their first language. The national education system has worked very hard to provide bilingual (Spanish & Mayan) education for the first three years of school but because some of the over 20 different Mayan languages are spoken by only hundreds of people, the quality of education is quite unequal. Most of the Mayan population live in rural areas where the schools have very few educational resources compared to those in the urban areas thus the learning opportunities for those rural students are very limited. Also, the educational system has serious efficiency problems. About 34% of the students fail their first year and only 42% of the students finish elementary school (essentially grades 1 to 6); less than 10% finish secondary school.

The national content standards recently developed were designed to serve several purposes, including establishing clear content and performance goals for each grade and standardizing the quality of education. Since 2009, the national assessments are administered by MINEDUC annually near the end of the school year in grades 1, 3, 6, 9 and 12, in Mathematics and Language Arts.

Operational forms have been developed through a strong common-item linking design to facilitate equating across years. The tests, linking, equating, and now standard setting, have all been supported through the use of Rasch scaling. A technical manual was developed during the assessment design and development process and was used as a guide to evaluate the degree to which each step was consistent with the *Standards for Educational and Psychological Testing* (AERA, APA, NCME, 1999).

Because standard setting, as a conceptual framework and a process, was new to Guatemala, the USAID and MINEDUC staff decided to investigate the methodology during the first implementation, with the idea that much could be learned through which a (potentially) more refined methodology could be used for final standard setting once the assessments have been finalized. It was also important to provide validity-related evidence regarding the appropriateness and feasibility of employing any given standard setting method in the Guatemala context. Through extensive review of the literature and guidance from external experts, the teams decided the Bookmark method would be most appropriate method, well suited to setting standards on multiple-choice tests, and one that they could implement with sufficient fidelity. An external consultant with extensive standard setting experience was obtained to help facilitate the standard setting process.

A paper summarizing the results of a randomized simultaneous replication of the standard setting process was presented at the 2009 NCME annual meeting (Rodriguez, Rubio, & Rego, 2009). This paper summarized the results of three independent panels setting standards through the Bookmark method in language arts (with 47 panelists in total) and three independent panels setting standard sin mathematics (with 51 panelists in total). In this report, several graphical displays were explored to understand differences in panels and their cut score recommendations and several statistical analyses were conducted to understand performance across the three rounds and across the three panels by subject area. In addition, within judge variation across rounds was examined. Here we would like to introduce judge (panelist) experience as a way of understanding the role of standard setting in Guatemala. We include some analyses of the panelists evaluative feedback immediately following the standard setting process and reflections from the facilitator of the standard setting sessions.

Design of the Standard Setting Replication Study

In the Language Arts portion of the study, 47 judges were randomly assigned to 3 independent panels (consisting of 17, 16, and 14 judges). In Mathematics, 51 judges were randomly assigned to three panels (consisting of 18, 17, and 16 judges). Groups were slightly uneven in membership because some judges failed to appear and participate in the process; judges were assigned to their panels prior to their arrival to move along the process. The panels were comprised of Guatemalan education stakeholders, including parents, classroom teachers, and school administrators. The Language Arts Panels were 67% female; the Mathematics Panels were 57% female.

The facilitators were trained together and used the same materials and procedures with each panel. Thus, each panel and their results were completely independent at the person level, including the facilitator and panel members. The process and materials were equivalent in each panel, with the exception of the different subject tests (Mathematics or Language Arts) for each set of three panels.

Judges included grade-specific teachers from various regions of the country that were assigned to the test in their area of primary instructional responsibility (Mathematics or Language Arts). The Guatemala national assessment system includes four performance levels: Unsatisfactory (Insatisfactorio), Should Improve (Debe Mejorar), Satisfactory (Satisfactorio), and Excellent (Excelente). This required panelists to set three cut scores to separate the four performance levels. A standardized (consistent) approach to the Bookmark method was employed by each of the independent panels.

Evaluative Feedback from Judges in Guatemala

Our hope was to use information from the judges' evaluations of the sessions to explore variation in cut scores within and across panels. Unfortunately, this largely resulted in little to no significant relations between the outcomes of standard setting sessions and perceptions of the process. There were a handful of findings that were interesting, yet not unexpected. These are highlighted here. The evaluation questions are provided in the Appendix, translated into English.

To facilitate analysis of Evaluation reports from the judges, item sets were assessed through exploratory factor analysis using Principle Axis Factor extraction. The 5 items regarding the effectiveness of the session (Effective Session) overall were composed of a single factor, with coefficient alpha of .79. The 7 items regarding participants did not result in strong factors and was subsequently divided into 4 parts: a single item about prior knowledge of standard setting (Knew Process), two items regarding participant interest in the process (Participant Interest), a single item stating that the work during the sessions was difficult (Work Was Difficult), and three items regarding the active involvement of the participants (Participant Involvement). The 5 items regarding the quality of the organization and materials of the session (Session Organization) resolved into a single factor, with coefficient alpha of .73. The 7 items regarding the effectiveness of the facilitator (Effective Facilitator) resolved into a single factor, with coefficient alpha of .84.

The first question of the Evaluation Form was: "Do you believe that in the test booklet used to set cut scores, all, some, or none of the items were in order of difficulty, from easiest to most difficult?" Overall, 39% of panelists believed all items were in order of difficulty; 3% believed that none of the items were in order of difficulty (Table 1).

Table 1

Judge Beliefs about Item Order by Panel (Frequencies)

		Items in Order		
Subject Area	Panel	All Items	Some Items	None of the Items
Language Arts	L1	5	9	
	L2	5	12	
	L3	9	6	1
Mathematics	M1	4	13	
	M2	9	6	1
	M3	6	10	1

Based on correlations among panelist responses (Table 2), when judges believed the items in the ordered-item-booklet to be in order of difficulty, they tended to report the session to be more effective. Judges reported to be more interested and more involved when they viewed the session to be effective, well organized, with an effective facilitator. Participant interest in the activity and their level of active involvement were related (*r* = .32). Participants who were less interested in the process were more likely to report that the work was very difficult.

Table 2

Spearman Correlations among Judge Evaluation Responses (n = 94 to 98)

		Items in Order	Effective Session	Knew Process	Participant Interest	Work was Difficult	Participant Involvement	Session Organization
Effective Session	*r*	**.230***						
Knew about Process	*r*	-.148	.136					
Participant Interest	*r*	.043	**.434***	.111				
Work was very Difficult	*r*	-.077	-.107	.147	**-.219***			
Participant Involvement	*r*	.152	**.411***	**.269***	**.319***	-.121		
Session Organization	*r*	.124	**.562***	.135	**.361***	-.121	**.322***	
Effective Facilitator	*r*	.114	**.548***	.089	**.453***	-.171	**.326***	**.526***

**p*<.03.

At the Panel level (across the six panels) none of the panel cut score variance indices (within judge across rounds, within panels across rounds, and between judges variation) were related to panelist perceptions of the session, including the cut score levels themselves. The only correlation that approached significance was between level of participant involvement and the highest cut score at the Excellent performance level (*n*=6, *r* = .73, *p* = .10). None of the correlations among within panel changes in cut score variability from Round 1 to Round 3 and evaluation responses were significant.

Are there significant connections between participant perceptions of the process and their involvement (participant evaluative feedback) and cut score variability within and between panels? Yes, to some degree, participant experience and perceptions of the process may help us understand variation within and between panel results. To do a better job of assessing these relations, a larger number of panels is needed, but the trends found here appear to be reasonable and consistent with other findings in the study.

Reflections from a Session Facilitator

Several important aspects remain salient upon reflecting on the process. These reflections are based on the impression session facilitators received from participants during the training and standard setting process across the panels.

1. During the procedure everyone is learning and teaching. While sharing their point of view about achievement, participants are molding each other's perception of the world in which they teach.
2. Participant perceptions and understanding about standards deepens to a new level of understanding.
3. The understanding of the importance of teaching and the roll teachers play in the nation is enhanced. This common understanding enhancement seems to create an environment of compromise during the standard setting process, particularly between rounds.
4. The tests are seen in a new perspective. The national tests usually are seen as simple instruments that can be easily completed, with no special emphasis in validity. The standard setting process helps participants realize that if the questions in a test are not strong

indicators of the domain they cannot be meaningfully used in assessing achievement in a way to support decision making. [Author note: this is to say that *validity* is now becoming an important concept to communicate and help educators understand, and is being recognized as an important characteristic to investigate.]

5. The term "Cut score" gets a new meaning through the standard setting process, beyond the simple meanings of pass or fail. Usually everyone understands that the cut score describes the lowest edge of expected performance (minimal competency), or mean (average competency) and upper edge performance (good or excellent competency). However, participants come to realize that a strong operational definition of the cut score clarifies the goals of achievement (such that the cut score is set in an appropriate position), bringing relevant depth to the meaning of the label "cut score."
6. Teachers having the experience of sharing with other teachers from different cultures in the country expand their awareness of what is happening in the country, seeing education beyond their particular community or region. This brings a new kind of depth to understanding achievement and performance.

Facilitators also made several "post-hoc" observations in terms of "what did we learn?" from the process of engaging multiple panels in the process and two years of standard setting. Recall that initial cut scores were established following the first administration of the assessments, which led to changes in the content standards themselves, followed by changes to the assessments, and finally new performance level standards. Some of these reflections also go beyond what was experienced during the standard setting process itself, but are included as indicators of larger impressions of this new process and new arena of accountability in

Guatemala. Some of these comments reflect the challenges in setting “national” standards across a largely rural country; the goal was inclusion in the process.

1. Moral issues about good and bad were common. Is it good to raise the standard and fail more students? Are the students failing going to get better or worse with the failing designation? Are the standards really going to help improve educational outcomes and the education system? Should we lower the standard so most students have the chance to continue?
2. Bookmark procedure is easier than Modified Angoff to be understood.
3. Recognition with a diploma or certification was welcomed by everyone and everyone made notice of who signed the certificates. Signatures from people in important positions made the recognition more valuable. Example of authorities that were valued included the Deputy Minister of Education, the international consultant facilitating the process, the director of the national assessment unit. Written recognition had more currency and was more valued than verbal recognition and thanks.
4. Facilitators and process assistants that had an easy “I’m here to teach and learn attitude” worked better that the facilitators with an “I’m here to teach” attitude. It was clear that teacher involved in the panels needed a little help to change from teaching to learning mode and the facilitator attitude was an important factor.
5. People coming from far away were tired with the trip, so they needed special attention to keep them interested and in the task.
6. Teachers coming from far away to the main city (the capital) found that the trip was a good chance to do some shopping and... they got carried away, returning a little late after the breaks. Isolated locations far from shopping areas worked better.

7. Food was important. Heavy lunches with lot of meat made people tired. Light lunches with some meat and light drinks (like lemonade) worked better.
8. Involvement in standard setting procedures gives educators the opportunity have a global-impact view and perspective of assessment. It broadens experience and understanding in a direct and powerful manner.

Standard Setting in Honduras

To improve the quality of education in Honduras and attain the Education for All (EFA) goals, the Honduran Ministry of Education developed a new national curriculum (DCNB: Diseño Curricular Nacional Básico) which has been introduced in schools throughout the country since 2004. In support of the implementation of the DCNB in classrooms, the MIDEH/AIR Project (the Improving Student Achievement in Honduras project, funded through USAID and operated by the American Institutes of Research) defined the national education standards, and developed the corresponding pacing guides (where the standards are organized by grade, on a monthly basis) for distribution to teachers throughout the country. The standards were then used to develop diagnostic, monthly formative, and end of grade summative assessments. All of these materials are aligned to the text books in Spanish, Math, Natural Science, and Social Science, and in turn, the textbooks and materials are aligned to the curriculum (DCNB).

The new curriculum and the aligned education materials present Honduras with an opportunity for systemic reform. This opportunity is complemented by the commitment of international donors in support of attaining the EFA goals in Honduras. With this support, the Ministry of Education has reproduced the standards and pacing guides for all teachers in the country, as well as diagnostic and formative test booklets for every student in the country. There have also been resources for the administration of end-of-grade summative tests on a national scale in 2007, 2008, and 2010. The results of these assessments are being used by the Ministry of Education to design teacher training strategies which respond to the needs of the teachers as reflected in the test results. By successfully implementing this systemic reform, there is hope in Honduras for substantial improvement in the quality of education.

Towards a Culture of Assessment and Accountability

The educational materials aligned to the DCNB were purposefully designed as practical tools for teachers. In fact, teachers recognized as leaders from all parts of the country participated in workshops organized to develop the materials. The workshops were facilitated by national and international experts in standards-based assessment. The workshops had different goals, such as definition of education standards, test development and item writing. The teachers who participated represented specific grade levels and subject matters. By participating in these workshops, the teachers not only helped develop the materials, but they were trained in the process. These teachers also became effective protagonists of these new materials. Consequently, the response in the classroom from teachers and school principals to the introduction of the new curriculum and the aligned materials has been largely positive.

This standards-based reform then proceeded with the development of an assessment system of internal and external evaluation, with standardized diagnostic, formative, and end-of-grade tests aligned to the content standards. Up until that time, some teachers designed their own diagnostic and formative tests, with a number of implications: the quality of the tests varied, as did the content, which made it impossible to compare results, even between two class sections of the same school, let alone from one school to the next. With the standardized assessments, teachers are compelled to teach what is scheduled in the national pacing guides, and the results of the tests are comparable nationally. Teachers, school directors and Ministry authorities are in a position to analyze test results and make decisions using reliable data.

The integration in the national education system of a standards-based assessment reform program creates the foundation for a culture of accountability. The results are reliable, and teachers, students, parents, and education authorities are interested in the results and their use for analysis and decision making to improve the quality of education. The primary role of accountability in the education sector is geared to achieve educational goals, locally and nationally. That is why a need for unified evaluation framework across grades and subjects became an emergent need, which led to the final step in building national assessment system: setting performance standards.

In spite of persistent efforts to empower the Ministry of Education and position the authorities at the leadership of this curricular and assessment reform, those kinds of innovations have been typically associated with the MIDEH/AIR Project. USAID and Project personnel have taken measures to minimize this attention to the Project, such as obtaining formal approval to reproduce the standards manuals, pacing guides, and the different tests without logos or recognition of the Project or U.S. government support. There are numerous reasons justifying the importance of empowering the Honduran Ministry of Education. For one, projects have a limited lifespan, and MIDEH is certainly not an exception. The education reform supported by the Project must be an integral part of the Honduran classroom long after the Project has ended. It is also important to note that for purposes of broad-based endorsement of the reforms, standards-based assessment has been presented not as the agenda of a specific project but as an extension of systemic reform which has improved the quality of education in other countries for years. Examples of leadership in this reform, such as Dr. James Popham, and his work especially in formative assessment, are used in Honduras for strategic modeling and to lend international credibility to the work being done locally.

Development of Honduran Performance Standards for Measuring and Reporting Learning

The content standards in Spanish language and mathematics have been defined and broadly used for national guidance in classroom instruction through the pacing guides, which were immediately followed by implementation of diagnostic and formative assessments (according to an external validation study conducted in 2006-2007, 77% of the primary level teachers were using the pacing guides to orient their teaching plans). As the next step, it has become important to build a system for monitoring the effects of these interventions, to build a summative standards-based assessment system for evaluation of student learning outcomes. This step required the development and implementation of performance standards.

At initial stage we needed to introduce the performance standards concept and involve the personnel of the evaluation unit (DIGECE) of the Ministry of Education. In many ways the challenge was changing a historical culture of presenting assessment results on the basis of percent-correct scores. Even the EFA indicator related to improving student academic achievement was established on the basis of the national average of percent-correct scores in Spanish language and mathematics. This is how it has been understood and reported by Honduran authorities and the international community supporting the EFA initiative. Inevitably, the society in general has expected the presentation of the results on the basis of percent-correct scores, because that is how the education system has always presented testing results. However, a closer review of the EFA proposal shows that, in fact, the attainment of EFA all goals are expressed as a percentage of students that meet clearly defined criteria, which lends to the conclusion that a criterion based on academic achievement is to be measured on the basis of the percentage of students who achieve a targeted academic performance level.

There is a substantial question that challenges the traditional approach: What does a percent-correct score really mean? Is it a measure of test difficulty or true student performance? What does it mean that a 9th grade student received a 32% in math and a 63% in Spanish? What does it mean that a school average was 56% in a given subject? Do we really know what those students know and are able to do? Given these unknowns, standards based on percent-correct scores (e.g., 90-100 is excellent, 80-90 is very good, etc.) are consequently vague. Thus, we needed a more meaningful frame of reference for evaluation of student performance. The issue was also tackled from the perspective of usefulness of percent-correct scores to teachers and education authorities for making timely and informed policy decisions.

We started with delineation of the difference between content standards and performance standards and how they are functionally related. Content standards define **what** students need to learn; performance standards represent a framework for evaluating **how much** of these knowledge and skills students should have mastered.

The theoretical discussion about performance standards was complemented concurrently in practice by the design of the scoring practices to be used by teachers during the administration of diagnostic and formative assessments. In workshops with the Ministry’s evaluation personnel, four basic performance levels were established, and general definitions were developed for each of the levels. On the basis of this work, the tests and teachers’ manuals were designed with instructions for each teacher to categorize the students in one of the four performance levels according to the number of test items answered correctly. This provided teachers with opportunities to evaluate each student but also gauge the performance of all the students according to the performance levels. This presentation of the results was especially useful for a school director who could ascertain which teachers and grade levels were having difficulties getting students to the more proficient performance levels. It is interesting to note that the teachers and school directors responded positively to the incorporation of this scoring system in the diagnostic and formative assessments, which in turn served to convince some of the more skeptical or reticent Ministry authorities.

In developing performance standards for summative assessments we used a framework that was widely researched and proven in practice (Cizek, Bunch, & Koons, 2004). This framework conceptualizes *Setting Performance Standards* as composed of two major procedures: Defining Performance Levels and Setting Cut Scores.

Defining Performance Levels is a procedure for conceptualizing performance levels to be used for evaluation of learning outcomes: deciding about the number and purpose of performance levels, choosing Performance Level Labels (PLLs), and developing general and specific Performance Level Descriptors (PLDs). The procedure typically employs focus groups method to solicit the opinions and judgments of field experts. In Honduras, we gave a high importance to formalizing and documenting this stage so we could use the performance levels descriptions for interpretational and reporting purposes.

Setting Cut Scores is a procedure for establishing cut scores on the operational tests used for classification of student performance into predefined levels. There are many different procedures that can be used for determination of cut scores and they can be categorized in two broad groups: procedures based on judgments about test items and procedures based on judgments about people and their work (Zieky & Perie, 2004). In Honduras we opted to use one of the most common and research supported methods for setting cut scores of educational assessments: the *Angoff* procedure (Angoff, 1971; Cizek, Bunch, & Koons, 2004).

Within each of the standard setting events in Honduras (April 2008 for grades 1, 3, and 6; December 2008 for grades 2, 4, and 5; and in February 2011 for grades 7, 8, and 9) these two procedures were run consecutively within a 4-day session involving a group of panelists for each grade/subject. Each group included about 15 teachers, with two facilitators per subject area.

Panelists were teachers that were nationally representative regarding the geographical location of schools (18 departments), rural vs. urban school location, and gender and ethnic background. They were recruited using the following criteria:

1. Be a teacher of the subject matter (Mathematics and/or Spanish) in the three grades.
2. Have at least 3 years of teaching experience in one of three grades.
3. Have successfully completed at least one in-service training (preferably related to assessment, for example, item writing workshop, standards development, etc.).
4. Implementing the new national education standards in his/her teaching work.

Facilitators were content specialists from MIDEH and Ministry of Education who participated in standard setting training. ***Technical staff*** was information technology and psychometric specialists from MIDEH that were trained to provide a comprehensive technical support for the procedure.

Setting Performance Levels

Performance levels refer to the degree of student mastery of content standards. The definition of performance levels is a process through which Ministry authorities and teachers, who are considered the classroom experts, guided by psychometric expertise, define Performance Level Labels, and general and specific Performance Level Descriptors. In

Honduras, workshops were organized in April 2008, December 2009, and February 2011, with teachers representing all 18 departments, as well as MOE representatives. During these workshops the performance level descriptors for Spanish language and mathematics for grades 1-9 were developed. The overall process of defining performance levels can be depicted by the following flow-chart:

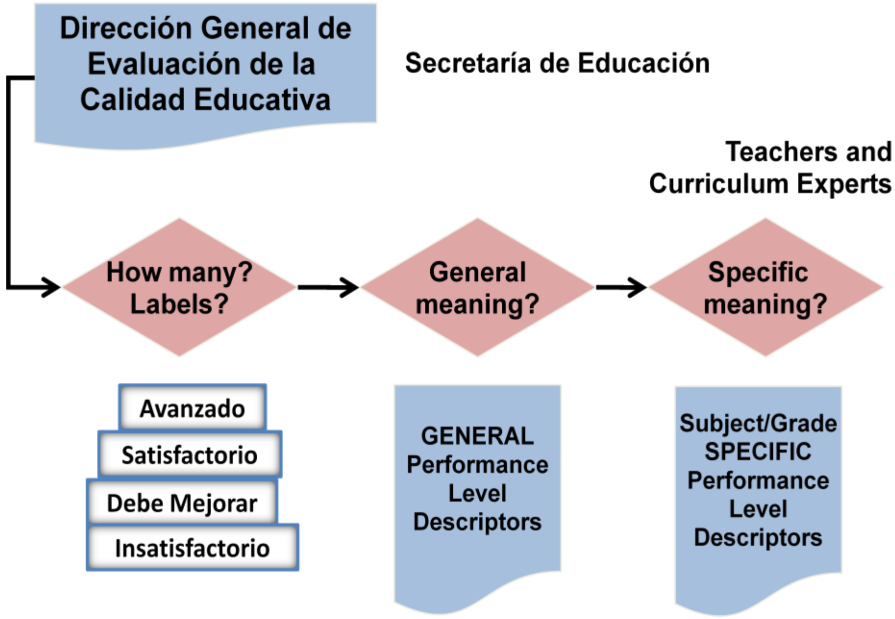

Dirección General de Evaluación de la Calidad Educativa
Secretaría de Educación
Teachers and Curriculum Experts
How many? Labels?
General meaning?
Specific meaning?
Avanzado
Satisfactorio
Debe Mejorar
Insatisfactorio
GENERAL Performance Level Descriptors
Subject/Grade SPECIFIC Performance Level Descriptors

Figure 1: Flow chart of process for developing definitions of performance levels.

The initial task was to determine and define the general performance level descriptors, with the following results, which have been deemed official by the Ministry of Education:

Unsatisfactory: *The student demonstrates limited knowledge of the content standards, and difficulties in solving simple problems. The student demonstrates a performance below the minimum acceptable mastery of the content standards.*

Needs improvement: *The student demonstrates partial knowledge of the content standards and the ability to solve easy problems. The student demonstrates a minimum acceptable performance in the mastery of content standards.*

Satisfactory: *The student demonstrates mastery of the content standards and the ability to solve a variety of problems. The student demonstrates an acceptable performance in the mastery of content standards.*

Advanced: *The student demonstrates a high level of mastery of content standards and the ability to solve problems with high level of difficulty. The student demonstrates superior performance in the mastery of content standards.*

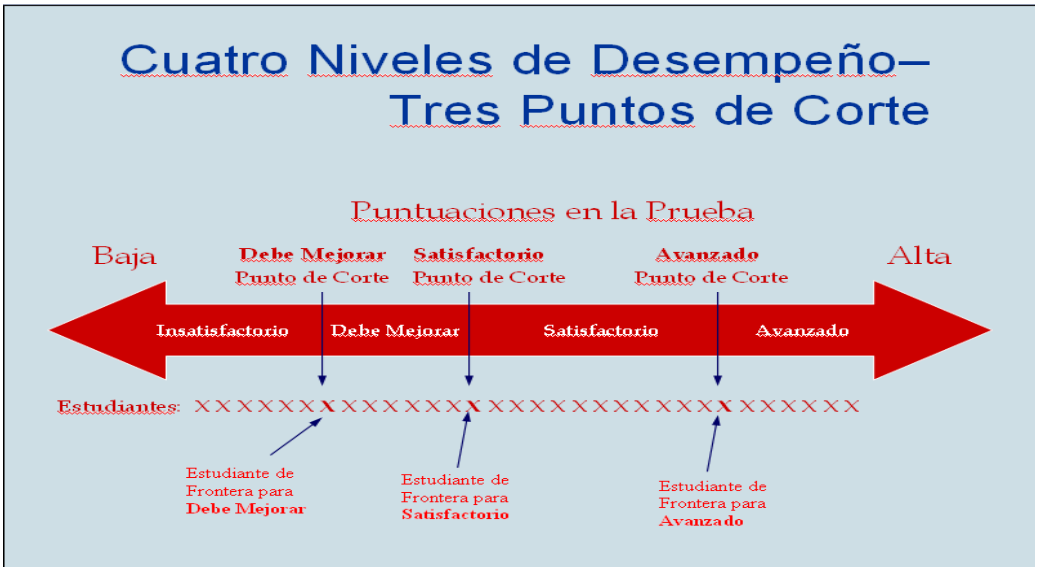
The next step was to define the specific PLDs, which describe what students should know and be able to do within each level of performance for each grade and subject separately. The typical procedure entails focus groups to solicit the opinions of all the participants, building to consensus through several rounds of discussions and analysis. The developed specifications will be used in score reporting and may also reflect on what to teach students, so resources are sought and methodologies implemented to help improve student performance, which in turn will maximize results and thus the goal of improving the quality of education.

Teachers, parents, and the public in general should understand the meaning of students' scores and what students that are classified into different levels of performance should know and be able to do. But more important is to identify the areas in which there is a need to improve the quality of Honduran education.

Setting Cut Scores

The Angoff procedure uses items from the test to elicit the experts' judgments about performance of students that are "borderline students" or "minimally qualified students" for each performance level. In the original Angoff procedure the judgments are given dichotomously (Yes or No) to answer the question: Should a minimally qualified student for particular performance level answer the item correctly? We used the modified Angoff procedure, which solicits the judgments indicating the probability that borderline students will answer the item correctly. The session material is composed of the item booklets and specially designed answer sheets for entering participants' judgments.

The borderline student is one who knows and can do everything that is on the lower level but does not necessarily know or can do what is in the upper level. For example, the student who is borderline to the "satisfactory" level knows and can do all that has been defined for the level "needs improvement", but he/she doesn't know or cannot do all that specify performance standards for the "satisfactory" level, rather, he/she achieved only minimally qualifying knowledge and skills for that level. The following figure illustrates this concept:



Cuatro Niveles de Desempeño–
Tres Puntos de Corte

Puntuaciones en la Prueba

Baja

Debe Mejorar
Punto de Corte

Satisfactorio
Punto de Corte

Avanzado
Punto de Corte

Alta

Insatisfactorio

Debe Mejorar

Satisfactorio

Avanzado

Estudiantes: XXXXXX**X**XXXXXXX**X**XXXXXXXXXXX**X**XXXXXX

Estudiante de
Frontera para
Debe Mejorar

Estudiante de
Frontera para
Satisfactorio

Estudiante de
Frontera para
Avanzado

Figure 2: Concept of borderline students and four levels of performance.

The Angoff procedure typically runs in three rounds. Each round is designed to foster increased consensus among panelists, although reaching consensus is not necessary to set the standards. Following each round of ratings, feedback is provided to illustrate the *agreement between panelists* and *impact data* that show how cut cores derived from their judgments would influence the percentage of students classified in each performance level. Panelists are provided with information relating to their own cut scores and group cut scores to evaluate their own agreement with the rest of the group, as well as impact data to discuss the percent of students classified in each performance level to answer the question: *How this information reflects your experience about what actually happens in the classroom?* Shown in more detail, the steps used in Honduras to implement Angoff's method were:

1. Developing the concept of *borderline student*. The method requires panelists to discuss about student knowledge and skills that are borderline between performance levels.
2. Panelists make judgments for each item on the test based on the experience and knowledge they have on the subject. They are instructed to look at each item and answer the following question for each performance level (except "unsatisfactory"): Which percentage of borderline students for the performance level "**needs improvement**" ("**satisfactory**" and "**advanced**") should correctly answer this item?
3. The judgments made by panelists are analyzed and the *feedback* is generated for evaluation and discussion after each of the three rounds. The feedback included agreement data (depicting the degree to which panelists agreed in their judgments) and impact data (percentage of students that would be classified in each performance level if the proposed cut scores of the group would be accepted, total and separately for males and females).

At the conclusion of the third round, the overall results are presented and the final opportunity is offered to panelists for discussion and global moderation to recommend the final cutoffs. These final recommendations are then intended to be reviewed and approved by Ministry of Education, so that they can be used in generation of annual score reports.

The cut scores, as well as impact data presented, are based on the operational testing MIDEH 2007 for grades in 1, 3; on MIDEH 2008 for grades 2, 4, and 5; and on MIDEH 2010 for grades 7, 8, and 9. It is important to note that the cut scores that are established on particular test forms cannot be directly applied to new test forms in subsequent administration years but they require the implementation of *equating* process to enable comparability between different test forms. Therefore, the MIDEH 2010 reports for grades 1 through 6 are based on *equated cut scores*, to ensure that the same criteria are used for classification of students in performance levels regardless of the particular test form used. Thus, in addition to performance standards, equating is also a required part of the national assessment system that aims to monitor educational progress across multiple years.

2010 MIDEH Results by Percent-Correct and Performance Levels

The results of setting performance standards were used as a framework for reporting student performance at national, regional, and school levels. The results were presented as percentage correct scores as well as the percentage of students in each performance level. The two different reporting methods are presented on the same page to enable comparison and highlight the importance of performance level reports for analysis and timely and informative decision-making.

For example, in the Spanish language results, in the percentage correct scores, one might draw the conclusion that the results are not that bad. But upon review of the results on the basis of performance levels, one is drawn to the problems identified in third grade, and to the need for special attention to improved reading and writing instruction during the early years of primary school.

In mathematics, the situation is much more critical. Upon review of the percentage correct scores, one might draw the conclusion that students start out quite well in first grade. However, the performance levels report demonstrates clearly that even in first grade, 34% of the students are in the unsatisfactory and needs improvement levels, and that in fact, that is where the problems begin and then accentuate in the following years.

Conclusions

In Honduras, the alignment of content standards with the curriculum, textbooks, and other educational materials has made possible the development and implementation of a national, world-class standards-based assessment system. The assessment instruments serve to measure learning achievements at the beginning of the school year (diagnostic tests), on a monthly basis (formative tests), and at the end of the school year (summative tests).

Performance standards have been defined conceptually (performance level descriptors) and operationally (cut scores), and national, regional, and school test results are reported on the basis of the percentage of students in each of the performance levels. Certain level of resistance could have been observed regarding the change from the percent-correct score reporting system, but with intensive promotional activities through media and country-wide workshops, the new system has a perspective to be accepted and fully implemented. Some additional observations and comments of the Honduran specific educational context include the following:

- The curricular changes proposed through the implementation of standards-based assessment are in accordance with national policy reform and the Ministry of Education evaluation plan as outlined in SINECE (Sistema Nacional para la Evaluación de la Calidad Educativa), the national evaluation system.

- The incorporation of performance levels in all three assessments: diagnostic, formative, and summative tests, assuring systemic coherence and practical and useful reports has been readily accepted by teachers and school directors, while the response at the central government has been supportive but not as enthusiastic.
- The force of the implementation of this systemic reform comes from the classroom, where dedicated teachers and school directors are implementing the system with quantifiable impact on student learning and achievement.
- The more significant challenges in successful implementation have been with the authorities at the central level, for different reasons: high turnover of Ministers and their senior teams (6 different ministers since 2004); the distractions, disruptions and strikes called by union leaders; and the constant turnovers of technical personnel resulting from the interference of politics and politically motivated change in governance.
- The timing of this political reform, to coincide with the EFA initiative, has had important and positive implications. The funds provided by the international donors were instrumental for reproducing and distributing materials nationally to teachers, students, parents, and authorities. This support has also been vital for the administration of end-of-grade summative testing.
- The government of Honduras, through the Ministry of Education, has invested heavily in teacher training. External studies demonstrate that this training has had limited impact on student achievement. By presenting test results by performance levels at the departmental level, teacher training strategies will shift to decentralized efforts where school directors and teachers not only participate in designing and facilitating trainings focused on their specific

needs, but will also be held accountable in their schools for transferring the knowledge acquired through training to more effective teaching in the classroom.

- The results of the diagnostic and formative tests administered in the classrooms by teachers are also presented according to performance levels. These results not only assist the teacher in identifying students who have fallen behind and require attention, but they also identify those students in the advanced levels who can assist as peer tutors.

Standard Setting in Chile

In Chile the initiative to develop standards was initially part of a package of measures adopted by the Ministry of education in the period 2002-2006 in order to affect educational quality and equity. Prior to this, the results of national assessments of learning showed stagnation, which resulted in strong criticism of the educational policies of the 1990s. In 2000 the results of the national measurement of elementary-school level applied learning (SIMCE) in 1999 were published with the results of eighth grade Trends in International Mathematics and Science Study (TIMSS). Both results generated great impact on public opinion and in the Ministry of Education.

The press addressed this information with a wave of criticism regarding the 1990s reform, and confirmed the need to revise national educational policy. In this context, the policy for basic education, in order to have greater impact on learning outcomes was reoriented: more detailed curriculum guidelines from first to fourth grade were made under the motto “bring reform to the classroom” and a campaign was initiated to focus on reading, writing, and mathematics (LEM). Furthermore, NGOs and universities offered advice to the schools with the worst results in metropolitan regions, and various educational reform programs were geared towards implementing the curriculum.

This moment of inflection in the educational policy is the prelude to the process of defining standards. On the one hand, existing levels of performance that described various achievements of the students on the national exams operated in practice as standards. In the year 2000 the government publicly committed to reduce the proportion of students in the lower level of performance within a period of 5 years, equivalent to the duration of the existing administration. Also as part of the LEM campaign, the government developed a guidance document for the schools, which described and illustrated the "performances to achieve" at various grades in reading, writing, and mathematics.

During this process of evaluating existing educational assessments and new reforms, authorities of the time were becoming convinced that the next step had to be the development of standards, which were considered a staple of the new emphasis on learning outcomes. Among the main concerns that were presented to a newly organized Advisory committee of national and international experts were: How are standards best conceptualized and defined? What considerations should be borne in mind to implement standards? What do researchers recommend based on the experiences of other countries?

The approach taken to develop standards

The Education Ministry of Chile distinguishes between *content standards* and *performance standards*. The first is described as maps of progress of learning in the central axis of each curricular area, and the latter, so-called levels of achievement, include descriptions of levels of performance in the areas that are monitored by the SIMCE national measurement.

Both the maps of learning progressions and the levels of achievement of the SIMCE describe qualitatively the learning of students; in the first case the learning expected in certain milestones across the school careers, in the second performance shown in the grades assessed in national measurement. In the case of the levels of achievement or performance standards, they are operationalized quantitatively in a cut score, which corresponds to the minimum score that a student must obtain on the test, so that their performance is graded at a performance level.

In conceptual terms, the approach taken is based on three fundamental ideas (Forster, 2007): the idea of growth or progression of learning, the idea of monitoring the learning with an explicit reference, and the idea of a national system of assessment articulated and coherent. These methods are more fully described in the Chile report.

An articulated and coherent evaluation system

The expectation of developing performance and content standards aligned with each other and under a same approach is to provide a common framework for the external measurement system for monitoring achievement at the local level (regions, municipalities, and schools). In this way, it is expected to promote and facilitate the articulation of information collected at the local level with information provided by the national measurement, promoting the use of both in the design of plans to improve learning opportunities offered to students. The potential of an articulated and coherent system of internal and external evaluation is the synergy generated when assessing the same learning process through the evaluation of classroom and external measurement (National Research Council, 2003; OECD, 2005; Pellegrino, et al., 2001).

Technical-political definitions of the nature of standards

Along with determining the approach to setting standards, at the beginning of the process it was necessary to make a series of technical-political decisions that defined the kind of standards that were to be established. Critical decisions related to the level of demand and rigor inherent in the standards of content and performance; the number of performance standards that would be elaborated (for example a single standard of performance for each grade assessed gives rise to two categories of performance - achieved/not achieved). The approach adopted in Chile was to define expectations based upon the evidence accumulated through national and international evidence on the results achieved by students.

Also, it was necessary to define whether the standards would be demanding or if they would be minimum standards. Standards were being designed to serve a mobilizing purpose across the education system, so the idea of minimum standards was discarded. For maps of learning progressions, we chose to describe expectations of learning equivalent to the average international performance exhibited in international tests, levels generally reached by 20 percent of Chilean students at the time of defining standards. In other words, it was decided that standards should be demanding for the Chilean population.

One or more standards of performance?

While it was desirable to convey a message of fairness and set a single standard for all students of the degree evaluated by the national assessment (SIMCE), given the heterogeneity of achievements in the country, this unique standard would have probably been very challenging for some schools and demanding to others. We determined that this scenario was inappropriate if the standards are intended to produce an effect of moving the level of performance, since a standard too distant from what can be achieved in reality does not motivate as it is seen as unattainable. Conversely, if the standard is a weak demanding yardstick for schools (if for example all their students achieve it), it is unlikely to motivate educators to generate action for improvement. For this reason, it was decided to describe three levels of performance, including basic, intermediate and advanced, giving rise to two cut scores.

Until 2009 the official discourse maintained the presence of two standards for each grade assessed, pointing out that the goal was to increase the proportion of students at higher level. The level called "Advanced" amounted to fully achieving the appropriate level of the learning progression map, whereas the intermediate level was equivalent to being in the bottom of the achievement level of the progression map.

Purpose and target: A role for teachers

Another decision to be taken in the preparation of the standards was to establish its central purpose. Here the central choice was between the use for accountability versus the use for education reform. While these two options are not mutually exclusive, the emphasis on one or the other has implications for the development of standards. So for example, standards designed exclusively for accountability creates a focus on the external measurement and communication of results without dealing with generating a frame of reference to monitor learning within schools. To fulfill its purpose, it would give increased centrality to parents as recipients, which would have had an impact on the public presentation of standards and in the language used.

The option adopted was to tip the balance towards the use by teachers as a central target of the standards. The learning progression maps are intended to provide clarity about the direction of learning in each area, where standards should be descriptive, under the assumption that it is necessary that teachers understand the performances or achieved learning by students. From this perspective standards are a device to monitor the strengths and weaknesses of the learning of students.

The emphasis was placed on the learning progression maps and levels of achievement rather than "standards" to promote acceptance by teachers, since it was found that the term "standards" generated rejection as it was associated with "uniformity".

The importance of the use of standards and communication

Throughout the development of performance standards, a goal was to assure that they are understood and used by the educational community, especially teachers. For this purpose successive validations were examined to ensure receipt by teachers, resulting in the development of a common communication design for learning progression maps and levels of achievement, and a strategy of gradually reporting to the system on the standards including an exploration of various strategies of communication.

From the beginning the development of standards in Chile was closely linked to the national assessment (SIMCE) and motivated by the need to improve the communication of their results. In the absence of clear definitions, both schools and the general public tended to erroneously interpret the national average score as the standard, without any reference to the achievements that this average represented in terms of learning.

From the point of view of assessment development, the central challenges were:

- How to reconcile the need for comparable results with previous years with the necessary changes in the evidence from its alignment with standards?
- With what methodology do we define the proportion of students achieving each level of performance in the tests?
- How can we promote the use of information reported by SIMCE and make the new way of delivering results - effectively - more understandable?

Several challenges arose in determining the definition of the cut scores which would determine the proportion of students in each category. The first was to assess the standards of performance and the technical challenge in this process of transformation of the system of measurement in reference to standards. The cut score is the operational version of the standard of performance, and conversely, the standard of performance is the conceptual version of the cut score. In this way, standards of performance and evaluation are intrinsically linked. In the case of the SIMCE, Bookmark (see (Zieki, M. et al. 2006) was the selected procedure to encourage the "reasoned opinion" Jaeger (2009) described as the intent behind standard setting procedures to ensure appropriate classification of examinees.

Finally, redesign of the reporting of results and achievement of the purpose of communication of results was a challenge. SIMCE previously publicly reported average scores, making invisible the variability of performance within schools. To prepare for the first delivery of performance level results (in early 2007), we began to test models of reports for different audiences in 2006, including focus groups of teachers, administrators, and parents regarding their interpretations and preferences with different presentations.

Social validation processes

Social validation and communication of the learning progression maps and levels of achievement, SIMCE committed to the process beginning in 2003 with the revision of the first draft of progression maps by committees of teachers and university academics. At the end of 2004, the drafts of all maps were subjected to consultation with subject area experts, who spoke on the clarity, progression, the demand and the relevance of the descriptions developed for all axes. While interesting comments were received through these reviews, the main questions raised by teachers were related to the relations between maps and curriculum and the usefulness for instruction. In short, learning progression maps can have an important mass media function, but they also have a deeper use as criteria to observe learning which requires a working face-to-face interaction with teachers looking at student work produced from tasks that allow observation of different levels of performance.

Consultations with the Ministry of Education, academics, and measurement experts supported methods for communicating the definitions and ideas behind performance standards. This included the specific labels to use regarding each performance level, as well as the relevance of the questions from each test to be used as examples illustrating the achievement of each level. This determined for example, the need to include explanatory notes or change some terms that were not clear for teachers, resulting in stronger definitions and labels for each level.

Information was also collected on the potential contribution of the performance standards as a method of reporting results of the SIMCE, including perceptions about the use and impact of reporting performance levels for each school in the education system. This information was used to develop the explanations of performance levels in terms of subject matter achievement, including an additional chapter in the report of results to support the use of the information.

Final reflections

The development of standards is a process fraught with decisions from beginning to end. This is not a special feature of this process, however, it seems to be relevant to highlight this point because the literature usually refers to the macro decisions related to the approach and the measurement of the standards, with little reference to the multiplicity of micro decisions involved in the definition of standards: the titles, the type of writing, the type of layout, the opportunity of communications, the implementation of the processes of participation and consultation, among many others. All are fraught with complex meaning and may eventually impact decisions and results.

This feature complicates and slows the process and requires a high tolerance of criticism and the need to make new adjustments. These dispositions are essential in a process where each micro decision must be subject to review to ensure quality and consistency. For those who are starting a process, the recommendation is to start preparing a map of decisions, which identifies the essential macro decisions (found in the typical guidance for standard setting), and then consider the time and control mechanisms needed to protect the consistency of the micro decisions (which inevitably arise through the execution of the process and implementation of the results). This is critical to ensure coherence in the entire system.

Another aspect to be considered in the process relates to those who will become the development team. The decision to be taken in this respect regards the extent to which standards will be developed with internal, usually insufficient capabilities, or if an external expert team must be recruited. If the process is carried out with internal teams, it is important to set aside time to develop necessary skills and capacities and to create conditions so that the team is supported to accomplish the necessary tasks. In addition, it is recommended that teams combine curriculum and evaluation specialists. This is necessary because from the beginning it is important to consider the evaluative component of standards. The combination of the two eyes contributes to the attention to both the type of learning you want to highlight in the standards on the one hand, and the evidence to determine what is attainable at certain times of the school experience.

Another aspect to consider is key to the definition of the type of standards that will be developed and refers to the relationship being sought between the standards, the curriculum, and the national assessment system. The articulation of the various parts is crucial to ensure that teachers receive a coherent curriculum message that orients and enriches their practices, rather than reduce them to what is measured, which will always be a part of the curriculum.

This relationship is very different in countries with decentralized educational systems, such as the United States, where standards are ultimately the curriculum, and in countries with centralized educational systems that have a national curriculum where the specified contents are expected to be addressed each year. The recommendation that emerges from this point, is that the approach to developing standards should be conducted in the light of the curricular instruments in use by teachers, offering additional value complementary to what they already know and do.

Finally, it is necessary to point out that it is not enough to develop standards, but to attend to their formulation and how they are understood by teachers and are articulated with the assessment and the curriculum. “Implementation” (if it is that you can speak of implementation of standards), is not mechanical and is again fraught with decisions and demands of articulation.

A standards-based policy will require for example, defined goals and commitment to deadlines for achieving them both at the national level and of each school. Support strategies must be established and capacity-building undertaken for those who are reaching their goals, and definitions should be developed for the type of consequences or measures to be taken at various levels if the pace of improvement is not as expected. The design and implementation of a policy organized around learning goals for students requires joint collaboration of the interior of the Ministry and the different teams and agencies responsible for different parts of the system: the evaluation of the performance and the continuous training of teachers, the assessment of the learning of school children, the establishment and monitoring of goals at different levels of education, the monitoring and the provision of support to schools, among others.

Probably the most relevant definition of a standards-based policy is the balance achieved between pressure and support. Educational systems vary in the extent that accentuate demands and clarity of goals, on the one hand, and strategies of support, on the other. The most effective combination to achieve improvement and high performance is high expectations and high support, provided that these are also of high quality. In the United Kingdom, for example, policies that combine standards and pressure exerted through national assessments on the one hand with effective strategies in support of teachers, on the other, resulted in a substantive improvement of learning in the areas of reading and math between 2001 and 2007. National assessments contributed to a better understanding of the objectives of the national curriculum by teachers, which was complemented with intensive training in service programs.

Application of the Bookmark Method in Chile

One of the recommendations delivered by the *Commission for the development and use of the measurement of the quality of the education system* (SIMCE), convened by the Minister of Education Mr. Sergio Bitar, in 2003, was that SIMCE should be referred to developing national performance standards and that results must be delivered in a way to be meaningful for teachers. Because of this, as of 2004, performance standards based on the SIMCE test, called levels of achievement were developed. In 2006, the first reports of results were released relating to levels of achievement of 4th grade education in Mathematics and Reading. The report then included standards for evidence of understanding of the Natural Environment in 2007 and in 2008 included evidence of understanding of the Social and Cultural Environment. The performance standards described how a student in a given school grade-level should perform regarding the achievement of some level of achievement. The appropriate interpretations were expected to answer the questions: What should students know and be able to do at each achievement level? What must you know and be able to do to reach the advanced level of achievement?

Each level of achievement has a qualitative component and a quantitative component. The qualitative consists of a description of what must demonstrate regarding what a student knows and can do at each achievement level. These descriptions were developed by the team (SIMCE) taking into account the curriculum framework and evidence of previous tests, in order to develop challenging (but achievable) achievement levels.

The cut score contributes the quantitative component of the achievement levels, providing the minimum score one needs to obtain on a SIMCE test to convey information about the knowledge and skills achieved at that level. Developing cut scores is a process that requires a group of persons to give expert opinion to determine the appropriate minimum performance that must be achieved to attain each achievement level. This process is governed by a set of standardized procedures which are intended to make the process relatively objective.

SIMCE chose to use the *Bookmark* procedure after an exhaustive review of the various methodologies available to set cut scores, based on the conclusion that this method better fit the characteristics of the assessment system (type of questions used in the tests, model of measurement, etc.). In addition, the *Bookmark* methodology has been used and validated in many systems of measurement at the international level. For example, it has been used in the majority of the states in the USA.

The selection, design, implementation and evaluation of this methodology were advised by staff of the Educational Testing Service (ETS). This is a summary of the technical procedures used with the SIMCE to establish the cut scores in *Mathematics* and *Reading* in 2006, *Understanding of the Natural Environment* in 2007, and *Understanding of the Social and Cultural Environment* in 2008.

The Bookmark Process Employed in Chile

Due to the progressive implementation of the standards process, procedures to determine the cut scores for 4th grade performance were scheduled in a phased approach for the tests administered between 2006 and 2008. In all cases, the procedure was directed by SIMCE specialists according to the Protocol of approved work standardized and recommended by ETS. For each subject area, a Panel was convened consisting of various specialists, charged with the task of applying their expert judgment to arrive at a general recommendation of cut scores.

A large and representative group of participants was convened to carry out the task of establishing cut scores, verifying that each panel had the presence of classroom teachers; university academics; research specialists; specialists of religious congregations; and specialists of the Ministry of Education. In addition, there was strong considered for participation of teachers from the metropolitan region, the South and the North areas of the country, as well as teachers in subsidized municipalities and private institutions. All panels were shaped in their majority by classroom teachers. The panel of *Mathematics* had 28 members, *Reading* with 27, *Understanding of the Natural Environment* with 34, and U*nderstanding of the Social and Cultural Environment* with 34 participants.

Using the results of the tests, questions answered students are sorted into a booklet, in order of increasing difficulty. The methodology of IRT used by SIMCE to calibrate the items and score tests, provides a strong measure of item difficulty, which can be used to sort test questions in the ordered-item booklet to support the Bookmark methodology.

The first activity was to reach consensus on what knowledge and skills characterize a student who achieves the minimum performances for each achievement level. Participants reviewed individually, the ordered-item booklet, then independently deciding which questions would be able to be correctly answered by the student at the minimum performance for each achievement level. Then, each specialist placed a separator or "bookmark" on the first question that this minimally competent student would not be able to answer correctly. The specialist repeated this procedure in three rounds, including the opportunity to discuss with the members of the panel the summary feedback from each round.

Once the cut score recommendations were obtained for the total panel, a confidence interval was created which was presented to a Technical Committee assigned the task of defining the cut score for each test, ideally placing it within the range recommended by the specialists.

Final Thoughts

The three countries involved in this paper present a range of experiences regarding the assessment of student learning, but all have similar contexts. This provides a range of experience and successes. Guatemala, Honduras, and Chile illustrate the experiences of educational assessment in América Latina and pose a substantial range in experience and success.

Despite recent advances, educational systems in the Latin American and Caribbean region (LAC) continue to face serious shortcomings. In Central America in particular, the overall quality of education is limited and LAC students consistently score near the bottom on international test comparisons. The gap in literacy between LAC countries and other regions of the world has widened in the last ten years. The inferior quality of education impedes the ability of the region to move forward politically and economically. LAC lags behind its competitors educationally: Young workers in the region enter the labor force with fewer years of education than do workers in countries of similar incomes in Asia and the Middle East. Children in LAC countries attend school an average of 5.4 years. In some countries, only ten percent of students graduate from the sixth grade without repeating a grade. Many drop out of school all together. (USAID, 2009)

The contexts in which these countries have addressed the goals of national assessment of student learning, a limited history of formal assessment, has presented an immense challenge – the language of large-scale assessment is absent in the Ministry of Education, schools, communities, and families. A great amount of institutional capacity building has been required. These countries have experienced new found goal setting and decision making capacity within their respective Ministries of Education. This has required the creation of additional bureaucracy and regulation making procedures. The technical challenges have been particularly acute, especially given the limited access to professional training opportunities. Overall, local capacity building has been evident and assessment technological advancements are seen as programs move toward modern measurement theory and gain access to resources to support continue professional development in educational measurement and psychometrics.

References

AERA, APA, NCME. (1999). *Standards for educational and psychological testing*. Washington DC, American Educational Research Association.

Cizek, G.J., Bunch, M.B., & Koons, H. (2004). Setting performance standards: Contemporary methods. *Educational Measurement: Issues and Practice, 23*(4), 31-50.

Kane, M.T. (2001). So much remains the same: Conception and status of validation in standard setting. In G. Cizek (Ed.), *Setting performance standards: Concepts, methods and perspectives* (pp. 53-88). Mahwah, NJ: Erlbaum.

Karantonis, A., & Sireci, S.G. (2006). The bookmark standard-setting method: A literature review. *Educational Measurement: Issues and Practice, 25*(1), 4-12.

Bransford, J.D, Brown, A.L., and Cocking, R.R. (1999) "How people learn: brain, mind, experience and school", NAP, Washington DC.

Coll, C. and Martin, e. (2006). The validity of the curricular discussion. In: revista PRELAC N ° 3. pp 6-27

Presidential Advisory Council for the quality of education (2006). Final report. Santiago, Chile.

Cox, C. (2001). The curriculum of the future. In: Perspectives vol. 4, no. 2.

Cox, C. (2006). Political construction of curricular reform: the case of chile in the 1990s. Teaching staff. Journal of curriculum and training of teachers, 10, 1.

Forster, M. and Masters, G. (1996-2001). Developmental Assessment. Australia: ACER.

Forster, M. (2002). Standards of learning. Curriculum and evaluation unit. Transcript of the presentation of Margaret Forster in the Telefónica building. Santiago: MINEDUC.

Forster, M. (2007). The arguments in favour of progress in Chile maps. 9Th UKFIET International Conference on education and development, 11,12 and 13 September.

R.M. Jaeger (2008) Setting Performance Standards for National Board Assessments: to reprise on research and development in Ingvarson, l. & Hattie, j. Eds. Assessing Teachers for Professional Certification: the first decade of the national board for professional teaching standards, advances in program evaluation, volume 11, 211-229, Oxford: Elsevier Ltd.

Hansche, L.; Hambleton, R.; Mills, C.; Jaeger, r. and Redfield, D. (1998). Handbook for the Development of performance Standards: Meeting the requirements of Title i. prepared by the U.S. Department of Education and The Council of to Chief State School Officers. Maryland: Frost Associates.

Constitutional Organization Act on education (Act No. 18,962/1990).

Law General education (2009).

Linn, r. (2001). Reporting School Quality in Standards-Based Accountability Systems. CRESST Policy Brief 3, National Center for Research on Evaluation, Standards, and Testing.

Linn, r. and Herman, j. (1997). Assessment standards-driven: technical and political problems in the measurement of the progress of the school and students. Peru: Grade.

Masters, Geoff b. (2005). "Continuity and Growth: Key Considerations in Educational Improvement and Accountability." Monitoring Learning. Recovered in October 2010 at: http://research.Acer.edu.au/monitoring_learning/10

Ministry of education (2003) evaluation of learning for a quality education, Commission for the development and use of the system of measurement of the quality of education.

National Research Council, NRC (2003). Assessment in support of instruction and learning: Bridging the gap between large-scale and classroom assessment. Washington, D.C., The National Academy Press..

OECD (2002). Definition and selection of competencies (DESECO): Theoretical and conceptual foundations

OECD (2004) Chile: review of national policies on education, Ministry of education, Chile.

OECD (2005). School factors related to quality and equity, OECD (2004). What Makes School Systems Perform: seeing school systems through the prism of PISA).

OECD (2005) Formative Assessment: Improving Learning in Secondary Classrooms, Paris.

PISA 2006: Science Competencies for Tomorrow's World, volume 1, page 189.

Pellegrino, J.W., Chudowsky, N. Glasser r. (2001). Knowing What Students Know: the Science and Design of Educational Assessment, Board on Testing and Assessment, Center for Education, National Research Council, National Academy Press, Washington, DC.

Perkins, D (1992). The smart school. The training of the memory to the education of the mind. Gedisa editorial.

USAID (2009). *Education in Latin America and the Caribbean*. Retrieved online July 28, 2009, at http://www.usaid.gov/locations/latin_america_caribbean/issues/education_issue.html

Zieki, M. and Perie M. (2006) A first on setting Cut Scores on tests of educational achievement, New Jersey, ETS, Educational Testing Service.