RESEARCH

DESIGN AND DEVELOPMENT OF EXAMINATIONS OF EDUCATIONAL ACHIEVEMENT AND QUALITY

EDUARDO BACKHOFF / ANDRÉS SÁNCHEZ / MARGARITA PEÓN /
LUCÍA MONROY / MARÍA DE LOURDES TANAMACHI

The authors are assigned to the Dirección de Pruebas y Medición (Directorship of Tests and Measurements) of Instituto Nacional para la Evaluación de la Educación (National Institute for the Evaluation of Education—INEE): Andrés Sánchez is assistant director of Tests of Mathematics and the Natural Sciences; Margarita Peón is assistant director of Tests of Spanish and Social Sciences; Lucía Monroy and María de Lourdes Tanamachi are project leaders.
Eduardo Backhoff is the director of Tests and Measurements of INEE. José María Velasco núm. 101, Colonia San José Insurgentes, Delegación Benito Juárez, CP 03900, México, DF, E-MAIL: backhoff@inee.edu.mx

Abstract:
The article describes the initial use of a new generation of national tests by Mexico's National Institute for the Evaluation of Education (Instituto Nacional para la Evaluación de la Educación—INEE) among sample groups of students who are representative in national and state terms. The aim is to evaluate the quality of the national education system, based on achievement attained at the elementary, secondary, and high school levels. A description is given of the objectives and principles assumed by INEE in evaluating learning, as well as the framework of reference of examinations of educational achievement and quality. The characteristics of national tests are defined: criteria, alignment with the national curriculum and the design matrix. The process is detailed for designing, constructing, administering and validating these examinations. The conclusion discusses the need to have evaluative instruments of quality to comply with the goals of INEE and thus help to improve education in Mexico.

Key Words: achievement tests, educational quality, basic education, Mexico.

Until 2002, the Dirección General de Evaluación (General Directorship of Evaluation—DGE) of Mexico's Secretaría de Educación Pública (Secretariat of Public Education—SEP) was the only government organization in elementary and secondary education responsible for evaluating the Sistema Educativo Nacional (National Education System—SEN). Since its creation in the 1970s, the DGE has designed and operated countless evaluative instruments and programs to obtain educational results from the country's elementary and secondary schools (Velázquez, 2000; DGE, 2000).

Although the results of these programs were not made public, and not all had the desired impact, the DGE contributed to putting on the table of political decisions the topic of educational evaluation as an indispensable means for improving the quality of national education; a presence that was reinforced to a large degree by the impact of the results obtained on international evaluations of learning, such as the PISA (Program for International Student Assessment). As a consequence of the above, the federal government, in its Programa Nacional de Educación 2001-2006 (National Education Program 2001-2006—PRONAE) considered the creation of rigorous, reliable mechanisms of evaluation that operate independently from the authorities and allow the rendering of accounts to society (SEP, 2001).

PRONAE established two major goals: *a)* the creation of the Instituto Nacional para la Evaluación de la Educación (National Institute for the Evaluation de of Education—INEE) and *b)* the redefinition of the functions of the DGE. The former was conceived as an independent, autonomous organization having the primordial objective of discovering and explaining the quality of the SEN as a whole, as well as presenting the results of its evaluations to the educational authorities and to society in general. The DGE, on the other hand, was redefined as an organization of central administration, with the purpose of carrying out diagnostic and formative evaluations oriented to providing feedback on decisions involving individuals and institutions (DGE, 2004).

With the idea of evaluating the quality of education in Mexico, more than five years ago the DGE developed what is now known as the Pruebas de Estándares Nacionales (Tests of National Standards—PEN). These instruments were utilized from 1998 to 2003, on representative samples of students throughout Mexico. However, upon its creation, INEE was made responsible for these tests, a function that was partially assumed with the analysis of the results of the test administered—still by DGE—in June of 2003. These results were published in its first report, *La calidad de la educación básica en México* (INEE, 2003).

Following the same methodological system proposed by DGE with regard to structure and contents, the INEE prepared a new version of the PEN, which it administered in June of 2004 to a representative sample of students throughout Mexico. The results were published in its second annual report, *La calidad de la educación básica en México: resultados de evaluación educativa 2004* (INEE, 2004a).

The PEN were designed with the idea of evaluating cognitive skills more than knowledge. For this reason, various tests of reading comprehension and mathematics were developed for elementary and secondary students. For elementary school, the evaluations were constructed by school year (2nd, 3rd, 4th, 5th and 6th); for secondary school, a single test was given to students in all three grades.

Unfortunately, the design and construction of the PEN suffer from important technical deficiencies that detract from the usefulness of comparing the results over time; for this reason, the tendencies in student learning cannot be discovered. In previous publications (Backhoff y Martínez-Rizo, 2004 and Backhoff, 2005), we have indicated that the deficiencies include the following: *a)* The specifications of content are not well-defined; *b)* An express methodology was not designed for comparison; *c)* The anchor questions do not have the same location on the different versions of the test; and *d)* no pilot study was made of an evaluation before administering the test.

Due to the above, it is impossible to compare in reliable form the results of the two exams, as various sectors of society have requested (Martínez-Rizo, 2004). In addition to these deficiencies, the documentation on the tests' underlying frame of reference is limited, and it is not clear if the minimum processes to validate these instruments were completed.

Furthermore, the PEN were not designed strictly with the purpose of evaluating the content of plans and national programs of study. Therefore, little can be said about students' mastery of curriculum contents. As a result, limited information would be given to the educational authorities regarding the attainment of the purposes of the Mexican curriculum (Backhoff y Martínez-Rizo, 2004; Backhoff, 2005).

These limitations are important on taking into account that INEE'S purposes regarding the evaluation of learning include the following: *a)* constructing a general vision of what students learn as a result of their formal schooling, *b)* discovering the strong and weak point of students' learning in the most important subjects, and *c)* allowing comparisons of scholastic achievement as well as discovering tendencies over time.

Considering PEN's limitations in attaining these purposes, the Technical Board of INEE opted to prepare a General Plan for the Evaluation of Learning. This Plan contemplates the construction of a new generation of national tests, the Exámenes de la Calidad y el Logro Educativos (Examinations of Quality and Educational Achievement—EXCALE); preparation began in February of 2004, and the first exam date at the national level was programmed for June of 2005 (INEE, 2004b).

Based on the above, the objectives of this article are: *a)* to present the purposes and principles of INEE in relation to the evaluation of learning; *b)* to describe the frame of reference of the EXCALE; *c)* to describe in summarized form the process of design, construction, application and validation of these tests; and *d)* to discuss the importance of having robust instruments for evaluating the quality of the educational services offered in Mexico.

The procedures summarized in this article would require much more space for a detailed explanation and justification. The goal is to produce, after this general presentation, a series of articles to detail the principal conceptual and methodological aspects.

Purposes and Principles of the Evaluation of Learning

The educational coverage of the basic level in Mexico has been resolved gradually. For the past decade, interest has been centered on researching the scholastic population's effective learning, and defining the basic learning that forms part of students' repertoire.

The evaluation of learning carried out by the National Institute for the Evaluation of Education has the intent of providing the nation with general knowledge of the scholastic achievement levels attained by students in the basic subjects on the national curriculum; these subjects are defined as instrumental (Spanish and mathematics), and those that cover large areas of the curriculum (natural and social sciences). In each grade evaluated,[1] the ideal solution in accordance with the nation's curriculum is identified (INEE, 2004b).

The evaluation of learning, however, can be understood in many forms, depending on the purposes for which it is used, as well as its scope, the objects of evaluation, the methods used to evaluate, and the theoretical/methodological postures on which it is based. To clarify for the reader the starting points of INEE with respect to the evaluation of learning, the purposes of this evaluation and its basic principles will be defined below.

The principal purposes of the evaluation of learning at INEE are:

- To discover the academic achievement of students at the state and national level, as well as the most important factors of context that explain the differences in the sectors studied.
- To contribute to the knowledge of the scopes and limitations of the National Education System and thus to promote the quality levels of elementary and secondary education in our country.
- To issue contextualized value judgments that serve to support documented decision-making.
- To supplement existing evaluation processes that are developed by other organizations and national and international programs (such as DGE and PISA).
- To provide elements to enrich the rendering of accounts owed to Mexican society.

INEE has defined that the large-scale evaluation of learning it carries out must be based on the following principles:

- It must be considered an external evaluation of the nation's SEN.
- It must be of high quality and adhere to internationally recognized standards and practices.
- It must provide information that gives a valid and reliable image of SEN as a whole.
- It must ensure that evaluations respect the value of fairness, in particular with regard to gender, sociocultural capital and ethnic groups.
- It must be completed in a clear and transparent manner, with the joint participation of teachers and specialists.
- It must be academically and socially legitimate.
- It must provide elements that help to improve the quality of the nation's educational system.

Frame of Reference of EXCALE

The documentary support for the new generation of national tests is based on the knowledge accumulated during more than one hundred years of psychometrics, as well as the experience of

approximately forty-five years of administering international evaluations of academic achievement. The studies that most influenced the preparation of this frame of reference[2] were those related to: *a)* the fundamentals of classical test theory and item response theory; *b)* the experience and frames of reference of national tests from other countries for evaluating educational quality, especially from Spain and the United States; and *c)* the international tests at the vanguard, such as the International Program of Student Evaluation (OECD, 2003) and the Third International Mathematics and Science Study (Schmidt *et al.*, 1997).

The theoretical framework of EXCALE is defined basically by the principles of classical test theory (Nunnally and Bernstein, 1994) and item response theory (Hambleton, 1993). The first provides the principles for constructing and validating criterion-referenced tests aligned with the curriculum; and the second offers the basic principles for the calibration and scaling of educational tests. Specifically, the new generation of national tests adopted the following theoretical/methodological criteria:

1) Criterion-referenced
2) Aligned with curriculum
3) Matrix design of questions
4) Selected-response (although constructed-response can also be used in some cases)
5) Scale based on the item response theory
6) Defined achievement levels for the interpretation of results
7) Use of parameters from classical test theory and item response theory to evaluate the quality of EXCALE and contribute evidence of validity

The EXCALE are *criterion-referenced* tests because they are designed to discover with precision the student's degree of mastery of a set of specific content. Thus the referent for interpreting EXCALE results is the amount and type of material the student handles in the universe of evaluated content or test construct (Popham, 1990). These tests will help us to evaluate the possession of knowledge and scholastic skills; in contrast with the norm-referenced tests, which are used to order and select individuals, and put greater emphasis on the specific contents evaluated.

On the other hand, the EXCALE are *aligned with the curriculum* because they are prepared with an ad hoc methodology for evaluating, with great precision, the curriculum contents (whether called skills, knowledge, competence, etc.) as defined by the national plans and programs of study (Contreras, 2000). Such curriculum alignment implies, according to Nitko (1994): identifying the important results on the curriculum, associating the actions of evaluation with essential content, defining the complete curriculum mastery used to develop the exam, and specifying the results of learning established by the official curriculum. Mexico has a national curriculum, free textbooks, and the more or less uniform training of teachers—ideal conditions for this types of tests; in contrast with other countries (like the United States) that, in order to evaluate the national level of education, must dedicate themselves to the enormous task of generating standards of content and execution in each discipline that will be used for aligning test design.

They EXCALE are tests of a *matrix* type because they are designed to evaluate a large amount of curriculum content, without submitting students to long days of test-taking. This requires constructing a set of questions that cover the complete mastery of the curriculum content to be evaluated, and later to divide that set into subsets and distribute the subsets among students, so that each student answers only some of them (Deng, Ferris and Hombo, 2003; Van der Linden, Veldkamp and Carlson, 2004). The matrix model limits the number of questions that each student answers (thereby reducing the time of the evaluation), while permitting total coverage of the selected curriculum content among all

examinees. As a consequence of the above, the individual grading of students loses precision, while error of measurement increases; such is not the case of the aggregate results at the level of the state, type of school and social stratum, which are the focus of interest for INEE (Gaviria, 2005).

Since the EXCALE are large-scale tests, their questions are basically *selected-response*, with a multiple choice design. All the questions of this type contain four possible response options, of which one is correct; partially correct options are not used. However, there are also open or constructed response on the Spanish tests to evaluate the expression of written language.

The *calibration and scaling* of EXCALE scores is carried out according to the principles and assumptions of the item response theory (Hambleton, 1993), specifically by using the model of one parameter, better known as the Rasch model (Wright, 1996; Linacre, 2005). One of the fundamental principles of this theory is the "single dimensionality" of scales; i.e., test questions must be proven to correspond to a single dimension in order for this type of scaling to be adequate.

The *system of interpretation* of EXCALE is one of the basic elements of its validity, and the results are interpreted according to the achievement levels or standards that describe what students know or are able to do with their learning. Establishing such levels is done through the "item correspondence"; the best known is the bookmark, described by Lewis, Mitzel, Green and Patz (1999). As Jornet indicated (2005), the usefulness of the information produced by these tests depends on the way the learning results of the National Education System are reported, with the possibility of establishing precise guidelines for improvement.

Lastly, based on the principles of classical test theory and item response theory, the EXCALE should adhere to rigorous *quality standards*, including: *a)* the clear definition of use and coverage, *b)* the use of rigorous procedures for design and construction, *c)* the use of standardized procedures for administering the test, *d)* the clear interpretation of results, and *e)* the exhibition of evidence of validity and reliability. The *validity* of EXCALE must be centered especially on the premise that test scores show how much students know and can do with respect to the national curriculum (Ruiz-Primo y Jornet, 2004).

Process for Designing, Constructing, Administering and Validating EXCALE

It is important to mention that designing national tests of quality requires following rigorous standards and guidelines for instruments of educational evaluation (AERA, APA AND NCME, 1999; Martínez y cols., 2000). It is also important to adopt a robust methodology that has proven its validity for the intended purposes; such is the case of the model for constructing criterion-referenced tests aligned with the curriculum, in which a central aspect is the collegial work of specialists and teachers. Due to its benefits, the INEE adopted this model for EXCALE (see Nitko, 1994), adapting it to national needs.

The process was defined in seven phases and sixteen basic stages. Table 1 shows the process of design, construction and validation of the INEE's national tests. The table offers in detail the procedures and products expected from each stage, as well as the personnel external to INEE that participate in each stage. It is important to indicate that in general, the products of each stage serve as inputs for the following; thus the process of generating this type of tests considers in part the process of their validation (Contreras, 2000; Contreras, Backhoff y Larrazolo, 2003).

As shown in this table and chart 1 (at the end of the document), various specialists, groups of advisers, committees of experts and the technical personnel from the Directorship of Tests and Measurement participate throughout the process. Each stage of the process uses diverse procedures, including: *1)* the documentation of similar processes of constructing large-scale tests of learning, carried out by institutions of recognized international quality; *2)* training directed to the five committees of specialists and teachers who participate in the process; *3)* the elaboration and preparation of materials

for the work of the five committees; *4)* collegial work for making decisions of greatest importance (type of matrix design); and *5)* contracted individual work, which requires the experience and knowledge of specialists in a particular subject (such as the design of a sample).

A brief description is given below of each one of the seven phases and sixteen stages in the process of designing, constructing, administering and validating this new generation of tests.

TABLE I

Process of Designing, Constructing, Administering and Validating EXCALE

Phases	Stages	Outside Participants*	Procedures	Products
I. General planning**	1. Design of general plan of evaluation	● Technical board ● Advisers in measurement and validation	1. Documentation 2. Seminars 3. Collegial work 4. Contracted work	1. General plan for the evaluation of learning, with the EXCALE frame of reference 2. General procedures manual 3. Technical manual for matrix design 4. Technical manual for point scale and skill levels 5. Theoretical framework of validation of EXCALE
	2. Design and elaboration of questionnaires of context	● Specialists in questionnaire design ● Specialists in the evaluation of learning		6. Frame of reference of questionnaires of context 7. Questionnaires of context for students, teachers, and directors
	3. Design and development of information system***	● Specialists in databases		8. System of databases of questions 9. Document that describes the structure and functioning of the database
II. Structure of EXCALE	4. Design of tests	● Academic committees (one per test)	1. Documentation 2. Training 3. Preparation of materials 4. Collegial work	10. Technical manual for the design of national tests 11. Curricular graph of each test 12. Table of contents for each test
	5. Specification of questions	● Committees that prepare question specifications (one committee per test)		13. Technical manual for preparing specifications 14. Specifications of questions for each test 15. Two revisions of question specifications
III. Construction of EXCALE questions	6. Elaboration of questions	● Committees that construct questions (one committee per test)	1. Documentation 2. Training 3. Individual and collegial work	16. Technical manual for constructing questions 17. Three questions per specification 18. Two revisions per constructed question
	7. Validation of questions	● Committees of validation and bias (one committee per test)		19. Technical manual for the validation of questions 20. Two reports of validation for each questions
	8. Pilot questions and questionnaires of context	● State coordinators of evaluation	1. Sample for pilot study 2. Pre-edition of questions and printing of booklets 3. Training	21. Technical manual for pilot questions 22. Sample of population 23. Technical manual for editing questions 24. Test booklets and printed questionnaires of context 25. Database with results of pilot study
IV. Production of EXCALE	9. Selection of questions and integration of blocks and forms	● Measurement advisers	1. Documentation 2. Analysis of questions	26. Technical manual for the psychometric analysis of questions 27. Report on statistical estimators of questions 28. Blocks of questions 29. Structure of forms (combination of blocks)
	10. Edition, assembly and printing	● Contracted printer	1. Editing of booklets and context questionnaires 2. Printing of booklets for optic readers	30. Technical manual for editing questions and booklets 31. Printed questionnaires of context 32. Printed booklets of EXCALE

Phases	Stages	Outside Participants*	Procedures	Products
V. Administering of EXCALE	11. Sample of population	● Specialists in sampling	1. Documentation 2. Collegial work 3. Contracted work	33. Technical manual for selecting samples 34. Updated sample framework 35. Sample of population and design
	12. Administering of tests and inputting of results	● All state coordinators of evaluation ● Contracted personnel	1. Documentation 2. Distribution of booklets in entities	36. Technical manual for administering booklets and context questionnaires 37. Booklet packages distributed in states 38. Database with inputted results of evaluation
VI. Analysis and interpretation of results of EXCALE	13. Initial analysis of results	● Advisers in measurement	1. Seminars 2. Statistical analysis of results	39. Technical manual on analysis of questions 40. Technical report of psychometric behavior of questions
	14. Establishment of achievement levels	● Committees on achievement levels (two per test)	1. Documentation 2. Training 3. Collegial work	41. Technical manual on establishing achievement levels 42. Document with achievement levels and cutoff points for each test
	15. Elaboration of technical report on the results of learning	● Advisers on measurement and validation	1. Documentation 2. Training 3. Collegial work	43. Technical manual on report of results of learning 44. Technical report on results of learning associated with variables of context
VII. Validation of EXCALE	16. Studies of validity of processes and test results	● Advisers on validation	1. Documentation 2. Research	45. Frame of reference of studies of validity 46. Frame of reference of each test 47. Technical reports of validity studies 48. Publications on validity of tests

* Personnel from Directorship of Tests and Measurement intervene in all stages of the process.
** General phase for new generation of EXCALE tests.
*** Starting in stage 4, the information system will receive the information produced throughout the entire process.

Phase I: General Plan

The first phase of the process has the primary purpose of establishing a long-term testing plan. Therefore, of particular importance is the participation of the Technical Board and the outside advisers in measurement and validation; those responsible for defining the purposes, principles and conceptual referents on which the remaining phases in the process will depend. This phase consists of three stages:

In the first stage, the general plan for the evaluation of learning is designed. This plan will state: the frame of reference of EXCALE; the process of designing, constructing, administering and validating the tests; the matrix design; and the model of achievement levels. The theoretical framework of academic learning and the program for validating the interpretations of the new national tests are elaborated (INEE, 2004b).

In the second stage, the context questionnaires are designed and elaborated—directed to students, teachers and school directors. These questionnaires will be completed along with the tests to explain the obtained results of learning.

Lastly, during the third stage, an automatic information system is designed and developed to house, maintain and handle the database (curricular structure of the subject, table of contents of the test, specifications and charts of questions, results of validity, parameters of pilot questions, and others) relative to the various tests generated.

Phase II: Structure of EXCALE

Starting in this second phase, all the stages of the process are specific for each test. This phase has the purpose of designing and justifying the structure of the examination, and on the other hand, preparing the specifications for all of the questions on the test. The participants in this phase are specialists in the

curriculum, in teaching the discipline, authors of textbooks, representatives of associations, as well as active teachers from various educational modes and strata.

In this phase, the fourth and fifth stages of the process are established. In the fourth stage, the academic committee of each test, formed by approximately ten specialists, makes an exhaustive curricular analysis of the corresponding subject and grade, in order to generate a graph (double entry table) of the subject to explain the curricular structure and the essential contents of importance for evaluation. Based on the graph, a table of contents is prepared for the test along with a justification of the corresponding curriculum content. This table explains the thematic areas or components, the topics and subtopics that develop into the curriculum content and intellectual skills to be evaluated, as well as the type and number of questions to use.

In the fifth stage, the ten-person committee of specialists and teachers that prepares the specifications for each test's questions, defines and describes in a detailed manner the characteristics each question must have. In other words, the content to be evaluated, its location on the curriculum, its importance, the intellectual skill required from the student, and the question format (characteristics of form, background, writing, etc.), are indicated, so that this description can serve as a guide or pattern for similar and as possible, equivalent questions to be constructed. The number of specifications for each test is variable, since it depends on the curricular extent of the subject.

Phase III: Construction of Questions of EXCALE

The third phase of the process corresponds to the preparation of questions, validation and pilot study. The purpose is to produce high-quality questions for the test. This phase includes the sixth, seventh, and eighth stages of the process.

In the sixth stage, the members of the committee (approximately ten textbook authors and active teachers) that constructs the questions of each test, independently formulate three questions for each specification. These questions will be reviewed by other specialists according to the question review manual, with primary emphasis on question/specification congruency. The process is repeated until the questions are complete to the satisfaction of the reviewers.

The resulting questions are pre-edited, saved in the information system, and sent to the committee on validity and bias to begin the seventh stage of the process. It should be highlighted that this committee is formed by active teachers in the 32 states of Mexico who represent diverse strata and educational modes. This committee reviews each question in relation to its content (curricular pertinence, degree of difficulty), test writing (language used, syntax) and cultural and gender biases (characteristics of content and writing that favor or damage a social group). Each question is reviewed independently by two teachers, and in the event of discrepancy, by a group of eight teachers who must reach a consensus. If necessary, recommendations are made to improve the questions or their elimination is justified.

In the eighth stage, a pilot study is carried out on an intentional sample of approximately five thousand students per test, in order to discover the psychometric comportment of the questions and detect the problems students face upon answering them.[3] This pilot study rehearses the real conditions of administering the test, including the training of coordinators and grading supervisors, administering surveys to students, teachers and directors, and the complete logistics of the evaluative studies. This stage culminates with the reading of student responses and the conformation of a database with the results obtained.

Phase IV: Conformation of EXCALE

The purpose of this phase is to edit the booklets for each national test, with the information compiled in the two previous stages; the questions to be used on the national tests are selected and grouped in blocks, and the diverse forms of each test are defined.

To make this selection, the psychometric comportment of the questions (among diverse groups of students) is analyzed, and the teachers' observations are considered in terms of validity of content and absence of bias. The analysis considers the standards of criterion-referenced tests; i.e., giving preference to content over psychometric comportment.

In the tenth stage, different blocks of questions and diverse forms are made. The blocks are formed with a small number of questions that as a whole can be answered in fifteen minutes and that share certain characteristics that make them equivalent in terms of content, difficulty level and variance, as well as extension. Once the blocks of questions are conformed, they are combined in order to assemble different forms, so that all the blocks have an equal proportion and distribution, to the degree possible. The number of forms can vary from one test to another. Lastly, the questionnaire of student context is added to each form in order to have the booklets printed and conclude the phase.

Phase V: Administering EXCALE

The fifth phase of the process has the main purpose of administering the tests to a national sample of students, as well as the questionnaires of context to students, teachers and directors. The results are inputted into a database for their subsequent analysis.

During this phase, the sample of students is designed according to the evaluative studies planned; the personnel in charge of coordinating test administration in the states are trained; the booklets are distributed, and lastly, the national tests are administered to the selected samples.

The design of these samples is completed in the eleventh stage, with the selection of schools and students in Mexico's 32 states. The design depends on many factors, including: the available sample framework, the purposes of the study or studies to be carried out, the representativity of the sub-populations to evaluate (rural, urban, private, etc.), the confidence level of the sample, and the type of matrix design of the tests. Once the sample is designed in a random manner, it is validated with the information provided by the selected schools.

In the twelfth stage, the training process is completed for the coordinators and examiners (close to 5,000), the booklets are distributed in the 32 states, the tests are administered to the selected students, the context questionnaires are administered to teachers and school directors at the two educational levels, and lastly, the results of the tests and questionnaires are read and inputted in a database. It is pertinent to indicate that due to the matrix arrangement of the application, the characteristics of the resulting database cause a need for some statistical analyses to be carried out in unusual manners, and for others to be replaced by ad hoc analyses for this type of tests.

For the case of open-ended questions that use only a national sample without state representativity, such as the writing section of the Spanish test, a pair of judges jointly grade each question based on a grading rubric or protocol so that the results can be added to the corresponding database. This stage of the process is very delicate because diverse problems occur that can easily invalidate the results of the evaluation. Some of the actions to strengthen the procedure are to standardize the judges' criteria, and to carry out reliability studies.

Phase VI: Analysis and Interpretation of Results of EXCALE

This phase consists of three stages and has the final purpose of preparing the technical reports on the test results and context questionnaires; elements which will be the basis of the annual reports on

learning published by INEE. To reach this goal, it is necessary to make an initial analysis of results to establish students' achievement levels and include them on the corresponding technical reports.

Thus, in the thirteenth stage, the first statistical analyses are made of students as well as questions, with special emphasis on the psychometric comportment of questions.

In the fourteenth stage, the committee of educational authorities, curriculum specialists, textbook authors and active teachers that establishes the achievement levels of each test, defines students' categories and achievement levels according to: *1)* the execution expected of them in "theory", and *2)* the real results on the respective test. This information is used to define the cutoff points for each level of achievement; i.e., the minimum and maximum scores that correspond to each level of the grading scale.

Lastly, in the fifteenth stage, the technical reports on the evaluated students' results of learning are prepared, considering the context variables and learning opportunities described on the context questionnaires administered to students, teachers and directors.

Of special interest for INEE is discovering the achievement level of the various sub-populations sampled (states and strata and educational modes), as well as the contents of the national curriculum that the students master. It is not inappropriate to indicate that the information generated in this stage contributes significantly to INEE'S annual publication: *The Quality of Education in Mexico.*

Phase VII: Validation of EXCALE

This phase consists of a single stage and has the goal of contributing diverse information on test validity and the interpretations derived from use.

Although for practical purposes the sixteenth stage is at the end of the test construction process, it actually begins at the moment the design ends (fourth stage). We can see two moments from the validity studies of EXCALE: *a)* during the construction process itself, and *b)* after its termination.

In the first case, the studies are directed to evaluating the quality and congruency of each stage in the process, and have the purpose of verifying the process itself, as well as generating information for correcting the problems detected during test construction.

In the second case, the studies have the purpose of contributing evidence on the veracity and limitations of the interpretations generated as a result of a test's use. The purpose is twofold: on one hand, to legitimate academically the evaluations generated by INEE; and on the other hand, to begin a process to improve the tests based on the information and documentation of these studies. The process to validate a test never ends; the evidence simply accumulates with regard to the veracity or falseness of the interpretations.

Conclusions

For INEE, the fundamental objective of evaluating learning is clearly to discover students' academic performance at the state and national levels, as well as the learning opportunities and factors of context that explain performance. The purpose is to issue value judgments that support documented decision-making and contribute to rendering accounts to Mexican society on the status of national education (Executive Branch, 2002).

The results of the evaluations carried out by INEE are expected to influence the following educational settings: national and state policies, national curriculum, programs of study and textbooks, school administration and management, teacher training and updating, and the society's opinion (INEE, 2004b). To achieve this impact, INEE must have, on one hand, evaluative instruments and procedures that are theoretically and methodologically sound; and on the other hand, the assurance that the results of the evaluations as described to the public are valid and reliable, so that teachers as well as educational

authorities can make documented decisions to improve education, and that academics and society can exert pressure on the corresponding authorities so that steps for educational improvement are taken in an informed manner.

It is important to emphasize two characteristics of the methodology employed to design, construct, administer and validate EXCALE: *a)* collegial work, in which numerous specialists and active teachers participate; and *b)* the documentation that is generated throughout the entire process, which leaves evidence of its validity. Collegial work consists of the participation of five committees of approximately ten members each; the advice of two groups of experts, with three specialists in each group; the contracted work of approximately five professionals, and the specialized work of close to 25 technicians from the Directorship of Tests and Measurement. In other words, for the design, construction and validation of each EXCALE, almost one hundred specialists and teachers are required, without considering the approximately 900 individuals who intervene in the pilot study and the almost five thousand participants when the test is administered at the national level.

Three aspects of the documentation generated by this methodology must be emphasized. First, most of the products are elaborated in sequential form and each product serves as input for the following step. Second, the products generated in each stage are reviewed and analyzed by the following committee, and although this committee cannot modify the actions of the previous group, it must indicate and if necessary justify its observations, so that evidence is provided on the problems that must be solved in the future to improve the instrument. Third, the documentation produced at the end of the process is abundant and rich in information; close to 48 products are prepared for each test, of which approximately one-half is common to all of them (such as the technical manuals) and the remaining half is specific to each examination (such as the tables of content).

With this methodology, we believe that two major goals established by INEE with respect to the evaluation of learning can be reached:

1) To discover the academic achievement of students at the state and national levels, as well as the most important factors of context that determine this achievement.
2) To contribute to discovering the scope and limitation of the National Educational System, and thus the quality of elementary and secondary education in our country.

These goals can be reached thanks to the benefits of the model of examinations aligned with the curriculum. Such is the case of EXCALE: the curriculum is the basis on which the examination is constructed, and decisions on what and how to evaluate are determined by the results of learning established in the curriculum. The central point of this model is to be able to guarantee that the set of questions we call EXCALE represents the universe of content that we call national curriculum; of course, the key to this guarantee is human judgment, which is present throughout the entire process of designing and validating these examinations, as proposed by Nitko (1994).

To conclude, we must state that this article, which explains in a general manner the model of the new generation of tests, represents the first step in presenting to the public the most relevant aspects of the methodology employed. However, this text should be interpreted with the necessary reservations since the model of evaluation described here will most certainly be modified and strengthened as information is received on the results of its use.

INEE, in accordance with its principles, is committed to report its methods and results to society—the reason for publishing this article. This commitment is highly relevant if we consider that evaluation must be a documented, open process; and not as in the past, when national evaluations were considered "black boxes" to which very few people had access.

Notes

[1] The grades that will be evaluated are the final grades of elementary, secondary, and high school (3rd year of preschool, 6th year of elementary school, 3rd year of secondary school, and 3rd year of high school), as well as the 3rd year of elementary school as an intermediary grade.

[2] In educational achievement tests, a "frame of reference" is commonly mentioned instead of a theoretical framework, since the first is more comprehensive and includes, besides the theoretical postures, the test characteristics, the methodological model of constructing the tests, and the conceptual framework of the areas evaluated.

[3] For which the work of close to 900 examiners is required.

Bibliographical References

American Educational Research Association (AERA), American Psychological Association (APA) and National Council on Measurement in Education (NCME) (1999). *Standards for Educational and Psychological Testing*, Washington: American Psychological Association.

Backhoff, E. (2005). "La comparación entre entidades: alcances y limitaciones de los *rankings",* en *Memorias de las jornadas de evaluación educativa*, México: INEE.

Backhoff, E. y Martínez-Rizo, F. (2004). "Resultados de las pruebas de estándares nacionales 2003: elementos para la comparación entre entidades", en *Memoria 2004, sexto foro de evaluación educativa,* México: CENEVAL.

Contreras, L. A. (2000). *Desarrollo y pilotaje de un examen de español para la educación primaria en Baja California*, tesis para obtener el grado de maestro en Ciencias educativas, México: Universidad Autónoma de Baja California.

Contreras, L. A.; Backhoff, E. y Larrazolo, N. (2003). *Curso taller para la elaboración de exámenes criteriales: manual para el Comité diseñador del examen*, documento mimeografiado, Ensenada: Instituto de Investigación y Desarrollo Educativo-UABC.

Deng, H.; Ferris, J. y Hombo, C. (2003). *A Vertical Scheme of Building the Naep Booklets*, document presented at the annual meeting of NCME, Chicago.

DGE (2000). *Balance de las acciones emprendidas entre diciembre de 1994 y octubre de 2000*, documento recuperado el 24/05/2005 en: http://www.sep.gob.mx/work/appsite/dge/index.htm

DGE (2004). *La evaluación en la Secretearía de Educación Pública*, documento mimeografiado, México: SEP-DGE.

Gaviria, J.L. (2005). *Propuesta de diseño matricial para las pruebas de español y matemáticas del programa de pruebas nacionales del INEE*, documento mimeografiado, México: INEE.

Hambleton, R.K. (1993). "Principles and Selected Applications of Item Response Theory", in Linn (ed.), *Educational Measurement* (3rd ed.), New York: MacMillan Publishing Co., pp. 147-200.

INEE (2003). *La calidad de la educación básica en México*, México: INEE.

INEE (2004a). *La calidad de la educación básica en México: resultados de evaluación educativa 2004*, México: INEE.

INEE (2004b). *Plan general de evaluación del aprendizaje*, documento mimeografiado, México: INEE.

Jornet, J. (2005). *El modelo de determinación de estándares de los Exámenes de la Calidad y Logro Educativos (EXCALE) del INEE de México*, México: INEE.

Lewis, D. M.; Mitzel, H. C.; Green, D. R. and Patz, R. J. (1999). *The Bookmark Standard Setting Procedure*, Monterey, CA: McGraw-Hill.

Linacre, J. M. (2005). *Winsteps Rasch Measurement Computer Program*, Chicago: Winsteps.com.

Marínez-Rizo, F. (2004). "Comparabilidad de los resultados de las evaluaciones", en *Memorias de las jornadas de evaluación educativa,* México: INEE.

Martínez-Rizo, F. *et al.* (2000). *Estándares de calidad para instrumentos de evaluación educativa,* México: CENEVAL.

Nitko, A. (1994, July). *A Model for Developing Curriculum-driven Criterion-referenced and Norm-referenced National Examinations for Certification and Selection of Students*, paper presented at the international conference on educational assessment and measurement, of the South African association, ASSESA.

Nunnally, J. C. and Bernstein, I. H. (1994). *Psychometric Theory,* New York: Mc Graw-Hill.

OECD (2003). *The PISA 2003, Assessment Framework: Mathematics, Reading, Science and Problem Solving,* mimeographed document.

Poder Ejecutivo (2002). "Decreto de creación del Instituto Nacional para la Evaluación de la Educación", *Diario Oficial*, 08/08/2002, México, D. F.

Popham, J. (1990). *Modern Educational Measurement: A Practitioner's Perspective*, Englewood Cliffs, N J: Prentice-Hall.

Ruiz-Primo, M. A., Jornet, J. (2004). *Acerca de la validez de los Exámenes de la Calidad y el Logro Educativos (EXCALE),* México: INEE.

SEP (2001). *Programa Nacional de Educación 2001-2006*, México: SEP.

Schmidt, W. *et al.* (1997). *Many Visions, Many Aims*, Vol. 2: A Cross-national Investigation of Curricular Intentions in School Science, Dordrecht, Holland: Kluwe Academic Publishers.

Van der Linden, W.; Veldkamp, B. and Carlson, J. (2004). "Optimizing Balanced Incomplete Block Designs for Educational Assessments", *Applied Psychological Measurement*, 28(5), 317-331.

Velázquez, V. (2000). "Hacia una cultura de la evaluación", en SEP, *Memorias del quehacer educativo 1995-2000*, México: SEP.

Wrigth, B. D. (1996). "Local Dependency, Correlations and Principal Components", *Rasch Measurement Transactions,* 10, 3, 509-511.

Article Received: August 4, 2005
Ruling: November 9, 2005
Second Version: November 25, 2005
Accepted: January 19, 2006

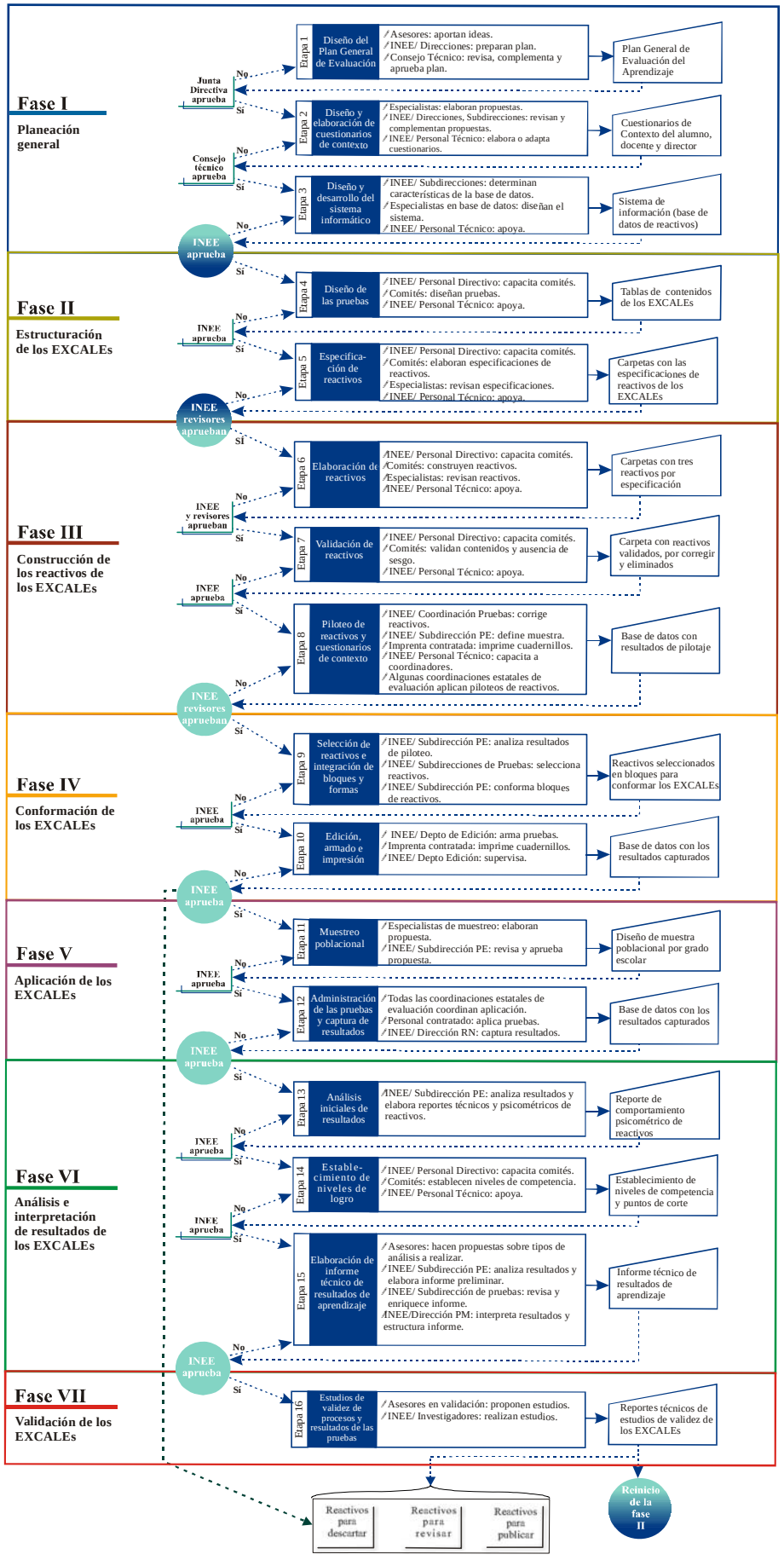
Fase I
Planeación general
Etapa 1
Diseño del Plan General de Evaluación
Asesores: aportan ideas.
INEE/ Direcciones: preparan plan.
Consejo Técnico: revisa, complementa y aprueba plan.
Plan General de Evaluación del Aprendizaje
Junta Directiva aprueba
No
Si
Etapa 2
Diseño y elaboración de cuestionarios de contexto
Especialistas: elaboran propuestas.
INEE/ Direcciones, Subdirecciones: revisan y complementan propuestas.
INEE/ Personal Técnico: elabora o adapta cuestionarios.
Cuestionarios de Contexto del alumno, docente y director
Consejo técnico aprueba
Etapa 3
Diseño y desarrollo del sistema informático
INEE/ Subdirecciones: determinan características de la base de datos.
Especialistas en base de datos: diseñan el sistema.
INEE/ Personal Técnico: apoya.
Sistema de información (base de datos de reactivos)
INEE aprueba
Fase II
Estructuración de los EXCALEs
Etapa 4
Diseño de las pruebas
INEE/ Personal Directivo: capacita comités.
Comités: diseñan pruebas.
INEE/ Personal Técnico: apoya.
Tablas de contenidos de los EXCALEs
INEE aprueba
Etapa 5
Especifica-ción de reactivos
INEE/ Personal Directivo: capacita comités.
Comités: elaboran especificaciones de reactivos.
Especialistas: revisan especificaciones.
INEE/ Personal Técnico: apoya.
Carpetas con las especificaciones de reactivos de los EXCALEs
INEE revisores aprueban
Fase III
Construcción de los reactivos de los EXCALEs
Etapa 6
Elaboración de reactivos
INEE/ Personal Directivo: capacita comités.
Comités: construyen reactivos.
Especialistas: revisan reactivos.
INEE/ Personal Técnico: apoya.
Carpetas con tres reactivos por especificación
INEE y revisores aprueban
Etapa 7
Validación de reactivos
INEE/ Personal Directivo: capacita comités.
Comités: validan contenidos y ausencia de sesgo.
INEE/ Personal Técnico: apoya.
Carpeta con reactivos validados, por corregir y eliminados
INEE aprueba
Etapa 8
Piloteo de reactivos y cuestionarios de contexto
INEE/ Coordinación Pruebas: corrige reactivos.
INEE/ Subdirección PE: define muestra.
Imprenta contratada: imprime cuadernillos.
INEE/ Personal Técnico: capacita a coordinadores.
Algunas coordinaciones estatales de evaluación aplican piloteos de reactivos.
Base de datos con resultados de pilotaje
INEE revisores aprueban
Fase IV
Conformación de los EXCALEs
Etapa 9
Selección de reactivos e integración de bloques y formas
INEE/ Subdirección PE: analiza resultados de piloteo.
INEE/ Subdirecciones de Pruebas: selecciona reactivos.
INEE/ Subdirección PE: conforma bloques de reactivos.
Reactivos seleccionados en bloques para conformar los EXCALEs
INEE aprueba
Etapa 10
Edición, armado e impresión
INEE/ Depto de Edición: arma pruebas.
Imprenta contratada: imprime cuadernillos.
INEE/ Depto Edición: supervisa.
Base de datos con los resultados capturados
INEE aprueba
Fase V
Aplicación de los EXCALEs
Etapa 11
Muestreo poblacional
Especialistas de muestreo: elaboran propuesta.
INEE/ Subdirección PE: revisa y aprueba propuesta.
Diseño de muestra poblacional por grado escolar
INEE aprueba
Etapa 12
Administración de las pruebas y captura de resultados
Todas las coordinaciones estatales de evaluación coordinan aplicación.
Personal contratado: aplica pruebas.
INEE/ Dirección RN: captura resultados.
Base de datos con los resultados capturados
INEE aprueba
Fase VI
Análisis e interpretación de resultados de los EXCALEs
Etapa 13
Análisis iniciales de resultados
INEE/ Subdirección PE: analiza resultados y elabora reportes técnicos y psicométricos de reactivos.
Reporte de comportamiento psicométrico de reactivos
INEE aprueba
Etapa 14
Estable-cimiento de niveles de logro
INEE/ Personal Directivo: capacita comités.
Comités: establecen niveles de competencia.
INEE/ Personal Técnico: apoya.
Establecimiento de niveles de competencia y puntos de corte
INEE aprueba
Etapa 15
Elaboración de informe técnico de resultados de aprendizaje
Asesores: hacen propuestas sobre tipos de análisis a realizar.
INEE/ Subdirección PE: analiza resultados y elabora informe preliminar.
INEE/ Subdirección de pruebas: revisa y enriquece informe.
INEE/Dirección PM: interpreta resultados y estructura informe.
Informe técnico de resultados de aprendizaje
INEE aprueba
Fase VII
Validación de los EXCALEs
Etapa 16
Estudios de validez de procesos y resultados de las pruebas
Asesores en validación: proponen estudios.
INEE/ Investigadores: realizan estudios.
Reportes técnicos de estudios de validez de los EXCALEs
Reactivos para descartar
Reactivos para revisar
Reactivos para publicar
Reinicio de la fase II

CHART 1

Diagram of the Design Process and Development of EXCALE *Tests*