TECHNICAL REPORT

Breaking Ground

Analysis of the Assessment System and Impact of Mexico's Teacher Incentive Program "Carrera Magisterial"

Lucrecia Santibáñez, José Felipe Martínez,
Ashlesha Datar, Patrick J. McEwan,
Claude Messan Setodji, Ricardo Basurto-Dávila

Prepared for the Mexican Ministry of Education

RAND EDUCATION

The research in this report was conducted by RAND Education, a unit of the RAND Corporation. This research was sponsored by the Mexican Ministry of Education (*Secretaría de Educación Pública*).

The RAND Corporation is a nonprofit research organization providing objective analysis and effective solutions that address the challenges facing the public and private sectors around the world. RAND's publications do not necessarily reflect the opinions of its research clients and sponsors.

Published 2007 by the RAND Corporation
1776 Main Street, P.O. Box 2138, Santa Monica, CA 90407-2138
1200 South Hayes Street, Arlington, VA 22202-5050
4570 Fifth Avenue, Suite 600, Pittsburgh, PA 15213-2665
RAND URL: http://www.rand.org/
To order RAND documents or to obtain additional information, contact
Distribution Services: Telephone: (310) 451-7002;
Fax: (310) 451-6915; Email: order@rand.org

PREFACE

Mexico's Carrera Magisterial is one of the pioneer teacher incentive programs in the world. It was instituted in 1992 and designed jointly by the federal and state education authorities and the teachers' union as a horizontal promotion system that rewards teachers with salary bonuses based on their performance. Teacher performance is evaluated based on a number of criteria, such as seniority, educational attainment, professional development, teacher performance, and student achievement. The program has never been formally and independently evaluated even though it assesses hundreds of thousands of teachers and is responsible for allocating millions of dollars in salary bonuses every year.

During conversations held in Mexico City in August 2003, Reyes Tamez Guerra, Public Education Secretary, and Jose Ma. Fraustro Siller, then Undersecretary of Educational Planning, both from the Mexican Ministry of Education, suggested that the RAND Corporation conduct an evaluation of the Carrera Magisterial program. The broad question they wanted addressed was: How can Carrera Magisterial be reformed to help it increase educational quality in Mexico? This monograph presents the results from RAND's work addressing this question.

While this report is concerned specifically with analyzing Carrera Magisterial's system of teacher evaluation, and is thus of particular interest to education policymakers at Mexico's Ministry of Education, it also offers general insights regarding teacher incentives and assessments that may serve to inform a broad audience of policymakers, educators, and the general public interested in the field of teacher reform.

This work was performed by RAND Education, a unit of the RAND Corporation. It was sponsored by the Mexican Ministry of Education (*Secretaría de Educación Pública,* or SEP). Any opinions, findings and conclusions, or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Mexican Ministry of Education.

CONTENTS

TABLES

FIGURES

SUMMARY

During the early 1990s, Mexico implemented a national education reform known as the National Agreement for the Modernization of Basic Education.[1] The federal government, state governments, and the national teachers' union (*Sindicato Nacional de Trabajadores de la Educación,* or SNTE) signed the agreement in May 1992. As part of this agreement, Mexico implemented Carrera Magisterial (CM), a program intended to give recognition to teachers and to provide economic incentives for superior teaching performance.

By using salary bonuses as an incentive tool, CM seeks to "help improve educational quality by rewarding and stimulating the work of the best teachers … and reinforcing teacher interest in professional development and continuous improvement" (*Comisión Nacional* SEP-SNTE, 1998).[2] CM is actually more akin to a salary adjustment or horizontal promotion system; i.e., it adjusts teachers' salary levels without changing their job descriptions. CM offers salary incentives to teachers who participate in professional development courses and who consent to be evaluated through teacher and student tests, as well as peer reviews. Most teachers in public primary and secondary schools are eligible to participate. Indeed, the majority of Mexican teachers participate in CM.

To allocate the salary bonuses, CM conducts annual evaluations in which teachers voluntarily participate. The CM evaluation system focuses on six dimensions, or "factors," as they are called in the program: highest degree earned, years of seniority, a professional performance rating by a committee composed of the teacher's peers, the teacher's score on a test following federal and state professional development courses, the teacher's score on a teacher knowledge test, and a score reflecting the teacher's classroom average on a standardized student achievement test.

After the evaluation concludes, each teacher is awarded a total point score. If this score is above a specified cutoff, the teacher is included in CM at one of five levels (A–E), receiving a salary bonus associated with that

[1] *Acuerdo Nacional para la Modernización de la Educación Básica* (ANMEB). Note that in Mexico, basic education includes three grades of preschool (declared mandatory in 2002; will be implemented gradually between 2004 and 2008), primary school (grades 1–6), and secondary school (grades 7–9).

[2] Authors' translation.

level. Teachers who enter the first level (A) receive a salary bonus of about 20 percent of their base salary (determined by the traditional salary schedule using seniority and education). Most of the teachers in CM are at this level. Teachers in the program's highest level (E) receive more than 200 percent over their salary base.

By implementing Carrera Magisterial, Mexico became one of the first countries in the world to link teacher salaries to performance in public schools. However, the political nature of the program—which is managed centrally by a joint commission of the Ministry of Education and the teachers' union (*Comisión Nacional* SEP-SNTE) — has resulted in ambiguities with respect to program objectives. The two main actors have not always agreed upon a unified vision. CM can be considered both a new salary schedule (SNTE's vision) and a pay-for-performance program (*Secretaría de Educación Pública's* vision).

THE PURPOSE OF THIS STUDY

Despite more than a decade of implementation at considerable cost to and effort of the Ministry of Education, CM has never been formally and independently evaluated.[3] Therefore, it is unclear whether the program functions adequately. Our study is not a comprehensive evaluation of the program as public policy. Rather it is an evaluation of its internal functioning and its potential for affecting educational quality. More specifically, our study evaluates the adequacy of the instruments used by CM to measure teacher performance. We also evaluate the effects of the salary incentives offered by CM on some indicators of educational quality. To do this, we focus on the following broad research questions:

1. Are the instruments used by CM to measure teacher and student performance technically sound? Are the test development and administration procedures adequate?

[3] This is not to say that CM has never been studied. For example, Schmelkes (2001) conducted an evaluation of the program using a sample of teachers working in marginal areas of the country. Ornelas (2002) summarized the program's main features and some of its effects. Other studies on the topic include García Manzano (2004), Tyler (1997), and Santizo (2002). None of these studies, however, have utilized the full dataset of the program to determine its impact at the national level nor have they specifically studied its evaluation system.

2. Are the instruments and procedures used to measure the peer-review factor adequate?
3. What is the relationship among the program's factors? Are they positively related to educational quality?
4. Is the program meeting its goal of helping to improve educational quality?

DATA AND METHODS

Data for this study come from four sources: the Carrera Magisterial program, the Ministry of Education's Evaluation Directorate, the National Program for Teacher Professional Development (PRONAP), and the National Statistical Institute (INEGI). We constructed a database with information on every teacher who participated in CM from 1998–2003, including scores on the six program factors and additional information such as teachers' gender, age, grade, and subject taught, and socioeconomic indicators of the school and region.

We employed various methodological approaches, mainly quantitative in nature. Whenever relevant, we complemented our analysis with targeted literature reviews to build on existing research and draw lessons from other countries.

To examine the reliability and technical soundness of CM's tests, we used standard psychometric indicators of internal consistency and item reliability. To examine the alignment of test content with the curriculum, we convened subject-matter expert consultants with knowledge of the education system in Mexico to evaluate subject-by-subject content frameworks. Last, to explore the levels of cognitive demand elicited by the teacher and student tests, we adapted frameworks from the literature that specified several cognitive categories, and the experts to make judgments on the cognitive demand of the items. To evaluate the impact of CM incentives on educational quality, we first performed an analysis of the relationships among the program's factors and explored the extent to which each factor was related to student test scores and to other indicators of educational quality, such as teacher test scores and peer review ratings. We then estimated two regression discontinuity models to understand (1) whether, for teachers attempting to enter CM Level A, the program's incentives have a positive effect on student test scores and (2) whether the salary incentives offered by CM have subsequent positive effects on student test scores after successful admission into CM or promotion to one of CM’s higher levels.

SCOPE OF THE ANALYSIS

Most of the empirical analyses in this study used data from 1998 through 2003. This time period covers five evaluation cycles, or *Etapas* as they are called in the program (*Etapas* 8 to 12). Before *Etapa* 8 (corresponding to the 1998–1999 school year), the program underwent considerable reforms.

We focused on primary and secondary classroom teachers, who in CM are referred to collectively as *primera vertiente*. The analysis included primary school teachers in grades one to six, as well as secondary school teachers teaching mathematics, Spanish, geography, history, and civics and ethics. We did not evaluate school administrators or academic support staff members, who are also eligible to participate in CM.

As described above, the CM evaluation system focuses on six factors: highest degree earned, years of seniority, a professional performance rating by a committee composed of the teacher’s peers, the teacher’s score on a test following federal and state professional development courses, the teacher’s score on a teacher knowledge test, and a score reflecting the teacher’s classroom average on a standardized student achievement test.

Our analysis of CM teacher assessment instruments was restricted to those used in grades 1 to 6 in primary school and the academic subjects listed above in secondary school. We did not evaluate the instruments used to test knowledge of the content provided in the various professional development courses, nor did we evaluate the content or procedures of the courses themselves. Our analysis of professional development was limited to studying the relationship between the scores obtained in this factor and other measures of teacher performance. In addition, our analyses of student test scores were restricted to teachers whose students are tested (grades 3 to 6 in primary schools and “academic” subjects—math, Spanish, geography, history, and civics and ethics—in secondary schools).

Our empirical strategy attempts to isolate the impact of CM on specific indicators of educational quality. The quantitative nature of the analysis does not allow us to identify the mechanisms or processes that might have caused any observed effects.

SUMMARY OF FINDINGS

Our analysis of the teacher tests found that, although some of the testing procedures are adequate, other features of the process are problematic. In general, the CM tests used to measure teacher knowledge have adequate or near adequate levels of internal consistency and reliability, represent broad coverage of the curriculum, and are well aligned with test guides and specifications given to teachers. However, the majority of the items in the subject-matter section of the tests, which measure teacher knowledge of his subject, demands only low-level cognitive responses. In addition, other test sections evaluating knowledge of pedagogy and Mexico´s education system have low levels of internal consistency reliability.

Although the student tests also represent broad coverage of the curriculum, other characteristics are faulty. These tests have lower levels of internal consistency reliability than do the teacher tests. In some cases, these levels are lower than those that would be desirable under international standards (particularly in the case of the student tests at the secondary level). As was the case with the teacher tests, the subject-matter-specific tests administered to secondary students demand mostly low-level cognitive responses.

Our evaluation of CM test development and administration procedures also produced mixed results. We found that testing procedures were generally adequate. However, there are no procedures in place to check for bias or ensure confidentiality of results on either the teacher preparation or the student achievement test. Although the teachers' professional preparation/knowledge tests appear to be carefully developed and monitored, the student achievement tests appear to suffer from less attention in their development, which in turn results in tests of comparatively lower quality. Moreover, we found a general lack of adequate documentation about item development and revision, and about technical quality, including reliability, particularly for the student tests. The documentation that is available reveals technical shortcomings in the statistical procedures used to develop and analyze tests.

Our analysis of the peer-review instruments and procedures found limitations as well. The instruments and evaluation criteria are not tied to explicit teaching standards on which to base subjective judgments about teaching practice. Furthermore, CM has not conducted empirical evaluations of the reliability of the peer review instrument or its ranking levels to ensure that they are based on meaningful definitions of quality and sound psychometric properties. In terms of the peer review procedures, we highlight concerns regarding purpose and process. Examples from the literature lead us to conclude that peer review is usually done with more formative purposes to help teachers improve practice. These examples also suggest that peer review is typically conducted by people outside the school or district to prevent conflicts of interest — protocol not followed within the CM program.

The results from our impact analysis of CM incentives on student test scores suggest that these incentives do not have any discernible effects on student test scores for primary school teachers and very modest positive effects on student test scores for secondary school teachers who are vying for admission into CM. In secondary schools, these effects are more evident after *Etapa* 10, when states adhered more closely to the minimum cutoff required for admission. Magnitudes of the effects for secondary school teachers range from 3 to 15 percent of a standard deviation, and they are only for teachers we classified as being within a group most likely to benefit from the incentive program (which was comprised of fewer than 20 percent of the total sample). After teachers were admitted into the program or received a promotion, we observed modest negative effects on their student test scores.

Because the composition of the sample varies from year to year, these results can be explained in part by attrition. They could also be the result of the CM teacher incentive structure. Perhaps, since salary bonuses are guaranteed for one’s entire career, teachers might exert lower levels of effort after receiving a promotion.

These findings are consistent with the literature on teacher incentives, which offers mixed evidence on the effects of salary incentives offered to teachers to improve student achievement. Most of the programs we reviewed in this study showed only modest improvements in student test scores, and these improvements (when they were observed) were often short-lived. Moreover, there was evidence that teachers devoted extra time to test preparation and this extra attention might have been partly responsible for the positive effects observed in some programs.

POLICY RECOMMENDATIONS

The mere fact that the *Secretaría de Educación Pública* (SEP) manages to annually administer more than 45 different tests to millions of students and hundreds of thousands of teachers with a limited budget is laudable. Perhaps the lack of sufficient resources is at the root of some of the deficiencies we identified in CM's assessment system.

The main recommendations of this study target the shortcomings observed in the teacher and students tests used to measure teacher performance. Even the best incentive systems might not show improvements in educational quality if the instruments used to measure them are flawed. It should also be noted that this research uses the program's own measures of quality as outcomes of interest (student test scores, teacher test scores, and peer review ratings). To the extent that these measures do not accurately reflect educational quality, our results will suffer from the same shortcoming.

Given our findings that the effect on student test scores of the incentives offered by the program is insignificant or very weak and that these very weak effects become insignificant after the bonus has been allocated, policymakers might consider reviewing the main features of the CM evaluation system, the factors it includes, and how these are assessed. To ensure that instruments used by test-based accountability systems are adequate, it is important to design and implement them according to internationally accepted standards of quality. To assure quality, technical manuals for test development and application and reporting of test results should be developed and continuously updated. These manuals should also include guidelines on reviewing items for bias. Moreover, they should incorporate detailed guidelines for confidentiality of the results and for communication and reporting of information to teachers, schools, state authorities, researchers, and the general public.

The CM teacher tests could benefit from several improvements. Currently, these tests do not increase in difficulty as teachers seek promotion into higher levels even though these levels are associated with higher salary bonuses. These tests are also much narrower than those used in other programs around the world, primarily measuring knowledge of one's subject matter. Other material on the test is of questionable value; the testing of knowledge of the Mexican education system is not supported by any theory of determinants of teacher performance. Policymakers should consider developing tests that measure teaching practice more precisely (using concrete measures of teaching competencies) as well as subject matter knowledge.

In addition to improving the teacher test, the peer review process should be further evaluated. International experience indicates that peer evaluation could be designed to give reliable and valid information about teacher performance while at the same time serving as an important formative feedback and improvement tool. The peer review factor could also be adjusted to avoid subjecting teachers to potential conflicts of interest by evaluating other teachers who work in the same building. A possible solution to the current conflict of interest situation is to invite teachers in the highest levels of CM to become mentors and peer reviewers. This last option, however, would depend on CM revising its assessment system so that teachers in the highest levels are those who have genuinely demonstrated greater teaching competencies.

A common element across the student and teacher tests and the peer evaluation instrument is the lack of standards that link the evaluation mechanisms to models of teaching or learning. We recommend developing performance standards to undergird all assessment measures used by CM. Standards should detail subject-matter content, abilities, and/or knowledge to be evaluated, accompanied by detailed performance criteria and indicators. They are invaluable tools in the development of testing instruments. For policymakers, setting standards represents an opportunity to clearly establish desired teacher characteristics and actions; for teachers, standards provide a guide or framework for improvement efforts; for evaluators, standards constitute basic specifications or terms of reference that serve to guide test development and testing procedures.

CM has placed a strong emphasis on professional development. But it appears that undergoing professional development at the national level is only very weakly related to improvements on teacher tests, student tests, and peer-review ratings. Because an evaluation of the professional development factor, its instruments, and its course content fell outside the scope of our work, such an evaluation warrants further consideration.

Other reforms of a more structural nature should also be considered. In general, incentive programs should avoid “double counting”; for example, if seniority and education are already determining the base salary, they should not also be used to determine the size of the bonus. Double counting is particularly problematic when these factors have been found to not be strongly linked to the outcomes of interest, as is the case in CM.

The program could also be modified so that it uses more than one year of achievement data when measuring teacher performance — using value-added methods that provide accurate and valid measures of teacher contributions to achievement gains over time. Such longitudinal assessments allow for providing continuous incentives for improvement as well as the possibility of implementing penalties or providing extra support if performance falls below acceptable levels. Evidence from the economics literature suggests that these kinds of actions could reduce "noise" in the performance measure as well as reduce risk to the worker, improving the efficiency of the incentives. It is also important that incentive programs minimize potential gaming behaviors, such as teaching to the test.

Teacher evaluation programs of the size and scope of CM benefit from technical and content advisory boards to supervise key psychometric and statistical aspects of its evaluations, curriculum coverage, and compliance with the necessary documentation on security, fraud detection, and other essential aspects of the assessment process. Such boards would work to ensure the quality of the tests and testing procedures.

All public policy programs strive to use resources efficiently. This is particularly important when, as is the case in Mexico, resources are scarce. One way to improve efficiency in the CM program would be to target bonuses only to those teachers who have strong incentives to perform.

Or, given that a national curriculum is already in place in Mexico, educational authorities could focus resources on developing a unified national testing system that could be used by CM as well as by other programs needing to measure student achievement. A unified national testing program would require tremendous technical, logistical, and administrative efforts. However, unified national tests would allow for external validation of the results of CM in relationship to other school reform programs at the federal and state levels. A unified system would also foster better use of available resources and technical capabilities. It would be important that policymakers analyze, with the assistance from technical experts, the option of consolidating the resources spent on CM's testing program as well as other testing programs currently administered by the Directorate of Evaluation of the Ministry of Education into one single testing program. Recent efforts by the Mexican Ministry of Education and other relevant organizations (such as the National Education Evaluation Institute) to develop a national education evaluation policy may represent a key opportunity.

ACKNOWLEDGMENTS

This research was made possible by the contributions of many individuals. We would like to especially thank Reyes Tamez Guerra and José Ma. Fraustro Siller of SEP in Mexico. Their vision and unwavering support at all stages of this project was instrumental to its successful completion.

We are indebted to Leonel Zúñiga Molina of *Unidad de Planeación y Evaluación de Políticas Educativas*, to Ana María Aceves Estrada of *Dirección General de Evaluación*, and to Juan de Dios Rodríguez Cantón of *Coordinación Nacional del Programa de Carrera Magisterial* for their help and advice throughout the project. Their perspectives and feedback were invaluable to our work. We would also like to thank Víctor F. Sastré Rodríguez, Mariano Díaz Gutiérrez, and Ernesto Ramírez Rodríguez of *Coordinación Nacional de Carrera Magisterial* and Sergio Martínez Dunstan, Joaquín Reyes Barrón, Proceso Silva Flores, and Carlos Goñi Carmona of *Dirección General de Evaluación* (DGE) for answering our numerous requests for information and helping us navigate the CM database. José Molina Luna, Jorge Alberto Díaz Guzmán, Jorge Rendón Albarrán, and Erika Rodríguez Vazquez of DGE helped us with data requests. We are grateful to Martha Lilia Cervantes C., of *Coordinación Administrativa* of DGE, who worked tirelessly on contracting aspects. Our appreciation also goes to Sonia Montes, Alejandro Rodríguez, Herguin Cuevas, Sasha de Silva, Leslie Serna, Alejandro Lara, and Beatriz Alcubierre, who served as consultants for the test analysis.

Several people at RAND contributed their knowledge and skills to this project. Charles Goldman served as senior advisor to the project and unofficial mentor. His support in this project was extremely valuable. Vi-Nhuan Le helped us at the initial stages of this project and undertook the analysis of the teacher tests. Lizeth Bejarano provided extensive administrative support and helped with document production. Maria Mendiburo helped us analyze the peer review literature. Ellen Tunkelrott, Sheila Logue, and Joanne Shelby helped us with the contractual agreements for the project. Christina Pitcher from RAND and Ana Cristina Enríquez from Columbia University edited this document and greatly improved its appearance.

We are extremely grateful to our internal reviewers, J. R. Lockwood, Sue Bodilly, and Laura Hamilton, who reviewed several drafts and provided

invaluable technical and general comments. Last, we are very grateful to our external reviewer in Mexico, María de Ibarrola Nicolín from Cinvestav. Her insights and knowledge about the education system and teachers in Mexico helped inform our findings and recommendations and greatly strengthened this study.

ACRONYMS

AERA	American Educational Research Association
ANMEB	*Acuerdo Nacional para la Modernización de la Educación Básica*
APA	American Psychological Association
CENEVAL (Ceneval)	*Centro Nacional de Evaluación para la Educación Superior*
CM	Carrera Magisterial
DGE	*Dirección General de Evaluación*
DGMME	*Dirección General de Métodos y Materiales Educativos*
ECIEE	*Estándares de Calidad para Instrumentos de Evaluación Educativa*
ETS	Educational Testing Service
IDANIS	*Instrumento de Diagnóstico de Alumnos de Nuevo Ingreso a Secundaria*
INEE	*Instituto Nacional para la Evaluación Educativa*
INEGI	*Instituto Nacional de Estadística, Geografía e Informática*
NAEP	National Assessment of Educational Progress
NCME	National Council on Measurement in Education
NRC	National Research Council
PARE	*Programa para Abatir el Rezago Educativo*
PRI	*Partido Revolucionario Institucional*
PRONAP	*Programa Nacional para la Actualización Permanente de los Maestros de Educación Básica en Servicio*
SEP	*Secretaría de Educación Pública*
SNED	*Sistema Nacional de Evaluación del Desempeño de los Establecimientos Educacionales*
SNTE	*Sindicato Nacional de Trabajadores de la Educación*
TIMSS	Third International Mathematics and Science Study

1. INTRODUCTION

More than a decade ago, Mexico implemented the National Agreement for the Modernization of Basic Education.[4] The federal government, state governments, and the national teachers' union (the *Sindicato Nacional de Trabajadores de la Educación,* or SNTE) signed the agreement in May 1992. The agreement had three main objectives: (1) reorganizing the educational system; (2) reforming curriculum and educational materials; and (3) reemphasizing the importance of teachers' roles in society (SEP, 1992).

The first objective was met through the implementation of a systemwide decentralization reform through a redistribution of the educational functions to the three levels of government: federal, state, and municipal. The second goal was addressed through the design of a new curriculum and the national textbooks.[5] The third objective was met through the retraining of teachers, improved compensation and benefits, support for affordable housing, and a program called "Carrera Magisterial." Carrera Magisterial (CM) was designed to recognize those in the teaching profession and provide economic incentives for superior performance.

The stated goal of CM is to "help improve the quality of education in Mexico through recognizing and supporting the work of the best teachers …" and "reinforcing interest in teacher professional development and continuous improvement (*Comisión Nacional* SEP-SNTE, 1998).[6]

CM is a horizontal promotion-type system; that is, it promotes teachers without changing their places in the hierarchy. Teachers in higher levels do not necessarily have management or other authority over teachers in lower levels, and teachers in lower levels do not report to teachers in higher levels. CM offers salary incentives to teachers who partake in professional development and consent to be evaluated through standardized tests of teacher and student knowledge.

[4] *Acuerdo Nacional para la Modernización de la Educación Básica* (ANMEB). Note that in Mexico, basic education includes primary (grades 1 to 6) and secondary (grades 7 to 9).

[5] Textbooks in Mexico are free and provided by the government for all public primary and secondary students.

6 Authors' translation from the original in Spanish.

The issue of the teachers' role in educational quality became particularly salient after Mexico's results on the Third International Mathematics and Science Study (TIMSS-95) were announced in 2001. The results were far from encouraging. Mexico placed either last or second to the last in most of the different grade level mathematics and science tests. Mexico also participated in UNESCO-OREALC's *Laboratorio Latinoamericano de la Educacion*, an international assessment that included only Latin American countries. This study found that Mexican student results in Spanish and mathematics were below the regional mean, and below other Latin American countries, including Argentina, Cuba, and Brazil (UNESCO-OREALC, 2002). Mexico also conducted its own internal assessments, administered by the Ministry of Education. The results of these assessments, summarized in the first report put forth by Mexico's Evaluation Institute, or *Instituto Nacional para la Evaluación Educativa* (INEE), revealed that close to two-thirds of the sixth graders did not achieve satisfactory competency levels in reading, while 87 percent did not achieve satisfactory levels in mathematics (INEE, 2003). These results spurred a debate in Mexico about the role of teachers in education, and about the impact, if any, that incentive programs like CM could have on teaching quality.

THE PURPOSE OF THIS STUDY

Despite more than a decade of implementation at considerable cost and effort to the Ministry of Education, CM has never been formally and independently evaluated.[7] Therefore, it is unclear whether the program functions adequately. This study is not a comprehensive evaluation of the program as public policy. Rather it is an evaluation of its internal functioning as well as an assessment of its potential for affecting educational quality. More specifically, this study evaluates the adequacy of the instruments used by CM to measure teacher performance. Secondly, it evaluates the impact of the salary incentives offered by CM on some indicators

[7] This is not the same as saying that CM has never been studied. For example, Schmelkes (2001) conducted an evaluation of the program using a sample of teachers working in marginal areas of the country. Ornelas (2002) summarized the program's main features and some of its effects. Other studies on the topic include García Manzano, (2004); Tyler and Esperanza (1997); and Santizo (2002). None of these studies, however, have utilized the full dataset of the program to determine its impact at the national level nor have they specifically studied its evaluation system.

of educational quality. To do this, the study focuses on the following broad research questions:

1. Are the instruments used by CM to measure teacher and student performance technically sound? Are the test development and administration procedures adequate?
2. Are the instruments and procedures used to measure the peer-review factor adequate?
3. What is the relationship among the program's factors? Are they positively related to educational quality?
4. Is the program meeting its goal of helping to improve educational quality?

These questions were derived after conversations between RAND and the *Secretaría de Educación Pública* (SEP) in Mexico. Answering the first two questions should shed light on the adequacy of the measures SEP uses for teacher quality (i.e., the tests administered to teachers and students and the measures used by peers and supervisors), as well as procedures used to administer the various tests. The third and fourth questions assess the program's impact on educational quality. Answers to these questions will result in policy recommendations to the Mexican Ministry of Education to improve CM.

EVALUATION FRAMEWORK

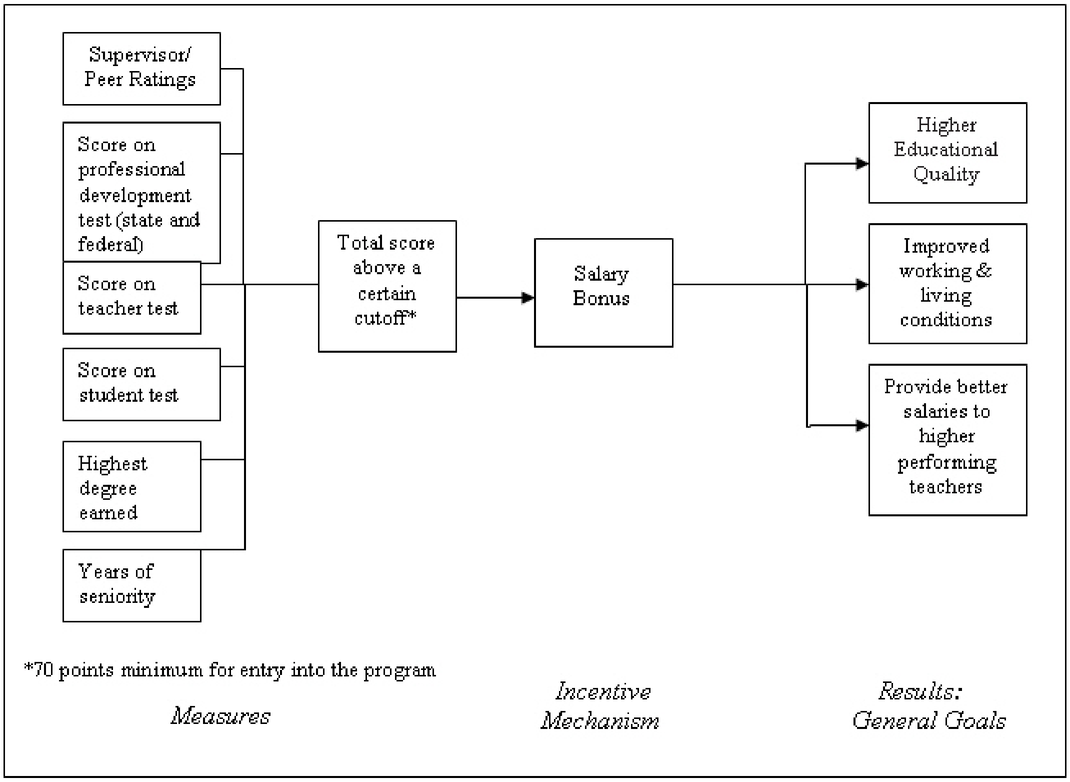
CM has three general goals, as stated in the program guidelines: (1) improve the quality of education in Mexico through recognizing and supporting teacher professionalism; (2) improve working, living, and social conditions of teachers; and (3) reward teachers for performance.

In addition, the program has five specific (or intermediate) goals: (1) strengthen the social appreciation of teachers; (2) encourage teachers to improve student achievement; (3) retain teachers; (4) reward and recognize teachers who work in low development areas and who teach special-needs students; and (5) strengthen interest in professional development courses (*Comisión Nacional* SEP-SNTE, 1998).

We use CM program goals as a starting point from which to build a conceptual map to guide our evaluation. Although we use all of the program's objectives to develop a conceptual map, this study focuses mostly on the first general objective and the second and fifth specific objectives.

Figure 1.1 describes the assumed causal relationships that lead to the aforementioned goals. These relationships underlie CM's design and assessment system. The measures correspond to the six factors (or dimensions of teacher performance and characteristics) evaluated by the program. The incentive mechanisms correspond to the rules to enter the program (in "program speak" this is referred to as an incorporation [first level] and promotion [subsequent levels]). The results depict the general objectives CM strives to meet.

Figure 1.1
Conceptual Map: Carrera Magisterial Teacher Assessment System and Program Objectives


Supervisor/ Peer Ratings
Score on professional development test (state and federal)
Score on teacher test
Score on student test
Highest degree earned
Years of seniority
Total score above a certain cutoff*
Salary Bonus
Higher Educational Quality
Improved working & living conditions
Provide better salaries to higher performing teachers
*70 points minimum for entry into the program
Measures
Incentive Mechanism
Results: General Goals

The relationships depicted in Figure 1.1 represent the following assumptions.

- teacher quality is mutable and can be affected by incentive provision
- individual monetary incentives are most productive (versus incentives at other levels, such as the school)

- recognizing and supporting teachers will lead to improved educational quality
- improvements in working and living conditions will lead to improved educational quality
- teacher quality is multidimensional (hence the six factors and their assumed relationship to educational quality)
- more professional development leads to higher educational quality
- improvements in working and living conditions happen mainly through salary bonuses
- a mix of objective and subjective assessments is appropriate to measure teacher performance
- student test scores are adequate measures of student knowledge
- teacher test scores are adequate measures of what a teacher knows

The research questions that guide this evaluation will test some of these assumptions.

DATA AND METHODS

Data for this study come from four sources: CM, the Ministry of Education's Directorate of Evaluation, the National Program for Teacher Professional Development (PRONAP), and the National Statistical Institute (INEGI). Using the data provided by all of these sources, we constructed a database with information for every teacher who participated in CM from 1998–2003 (program stages, or *Etapas* 8 to 12), that included highest degree earned, seniority, gender, grade or subject taught, scores on the six program factors, indicators of professional development, raw teacher and student test scores and socioeconomic indicators of the school and surrounding region.

The resulting database has five important limitations for our analysis. First, since CM only reports student test scores at the classroom level, individual students cannot be identified nor followed from one year to the next.[8] However, it is possible to link the average classroom test score to the teacher. Second, program participation is voluntary, hence the data only

[8] Due to budgetary restrictions, the student tests are administered in a "fragmented" form. Instead of applying a full exam with, for example, 100 items to all the students in the class, the exam is broken down into five different tests of 25 items each, and administered to the class. This means that individual results are not representative of the knowledge being tested by the entire exam.

include a self-selected group of people who choose to participate in CM. Third, teachers are in the database only in the years that they participate in CM.[9] Fourth, the database does not have any information on teachers who do not participate. And, fifth, there is no student achievement information for certain types of teachers including preschool teachers, first and second grade teachers, indigenous school teachers, secondary teachers in distance education (*telesecundaria*), and secondary teachers in technological and art subjects, because the program does not test their students.

Additional data were collected from CM and from the Ministry of Education to evaluate the technical quality of the instruments used by CM to test teachers and students. These data consisted of item-level data on student responses (available only cross-sectionally), content and curriculum standards, descriptive statistics on test attributes, and documentation on testing procedures. In addition, we conducted interviews with CM and DGE officials to fill in gaps where no written information was available.

METHODOLOGICAL APPROACH

This evaluation uses various methodological approaches, most of which are quantitative. Whenever relevant, we complement our analysis with literature reviews to build on existing research and draw lessons from other countries.

To examine the reliability and technical soundness of the CM tests, we use standard psychometric indicators for internal consistency and item reliability. In the case of the student tests, the availability of item-level data permitted a closer exploration of these issues using variance components and factor analysis. To examine the alignment of test content with the curriculum we used a qualitative approach that involved convening expert consultants with subject-matter expertise and knowledge of the education system in Mexico and creating subject-by-subject content coding frameworks. Last, to explore the levels of cognitive demand elicited by the teacher and student tests, we adapted frameworks from the literature that specified

[9] Even though the full evaluation can only be made up of the factor point scores obtained during the *Etapa* (*Comisión Nacional SEP-SNTE*, art. 6.3.5), teachers may take the evaluations every year if they wish. During their wait period (the years after a promotion that teachers must wait before becoming eligible for a subsequent promotion), teachers can be evaluated each year and choose the highest score obtained for the year when they become eligible for a promotion (see Chapter 3).

several cognitive categories and asked the experts to make judgments on the cognitive demand of the items.

Assessing whether CM's incentives are having a positive impact on educational quality is complicated because of the lack of an experimental design with a control group that would have allowed us to identify causal program effects. In our attempt to evaluate the impact of CM's incentives on educational quality, we first performed an analysis of the relationships among the program's factors. We then used a quasi-experimental regression discontinuity analysis, using student test scores as measures of educational quality. These analyses will help us determine whether the incentives provided by CM induced teachers to improve their test scores for those teachers attempting incorporation into CM; and, second, whether the salary incentives offered by CM had subsequent positive effects on student test scores after a successful incorporation or promotion into one of CM's higher levels.

LIMITATIONS TO THE STUDY AND SCOPE

One of CM's objectives is to improve educational quality, but CM's guidelines are vague about how this is measured or what is included in this concept. In this study we define educational quality as either student test scores, teacher test scores, or peer-review ratings.[10] To avoid confusion, we clearly state which measure of quality is being considered in each of our analyses.

The choice of these measures of educational quality was forced upon us by the nature of the data. Unfortunately, there are no national or other kinds of student or teacher quality assessments in Mexico that can be linked with the CM dataset. This makes it impossible for us to use external measures of educational quality to test the program's impact. This also affects our ability to determine whether CM's testing instruments are externally valid. The fact that we must rely on CM's own measures of quality limits our analysis to what CM's instruments can tell us about student achievement and teacher competence.

[10] In some cases we use CM-assigned point scores for the teacher or student tests. In other cases we use the raw scores (the test score expressed as percentage right answers on a scale of 1 to 100) obtained by the teacher *before* these scores are transformed by CM into program points.

Our evaluation focuses on primary and secondary classroom teachers, who in CM are referred to as *primera vertiente*. Because of the limited time span we had to perform the evaluation, we did not study school administrators or academic support staff (who are also eligible to participate in CM).

Our analysis of CM teacher assessment instruments was restricted to those used in grades 1 to 6 in primary schools and to those used for teachers of math, Spanish, geography, history, and civics and ethics in secondary schools. Our analyses of student test scores was restricted to teachers whose students are tested (grades 3 to 6 in primary schools and the academic subjects listed above in secondary schools).

We did not evaluate the instruments used to test for knowledge of the various professional development courses. The analysis of the professional development factor was limited to the relationship between teacher scores in the factor and other performance measures produced by the program.

Most of the empirical analyses in this study focus on years 1998–2003 (*Etapas* 8–12 of the program). We decided to start our analyses using data from 1998 because the program underwent considerable reforms in the 1997-98 school year.[11]

ORGANIZATION OF REPORT

This report is organized as follows. Chapter 2 presents a review on teacher incentive programs around the world. Chapter 3 provides a description of CM and teacher participation. Chapter 4 presents results from our analysis of CM tests and testing procedures; Chapter 5 analyzes CM's peer review instrument. Chapter 6 presents our results of the impact analysis of CM's salary incentives on student test scores as well as an analysis of the interrelationships among the program's factors. Chapter 7 describes our policy recommendations, some final considerations, and directions for future research.

[11] These reforms were largely undertaken to improve the peer-review factor. Prior to 1998, this factor had the greatest weight in the evaluation, and all teachers were receiving the maximum number of points. To resolve this situation, a "correction adjustment" was implemented (García Manzano, 2004). The reform decreased the weight of this factor to 10 points. See more about this in Chapter 3.

2. LITERATURE REVIEW: TEACHER INCENTIVE PROGRAMS IN THEORY AND PRACTICE

This chapter discusses findings from the literature on teacher incentive programs to consider the theory and assumptions that underlie teacher incentive programs in general and whether these assumptions are likely to hold in education. A review of recent international experiences with teacher incentive programs is intended to highlight the effects of such programs on student test scores and other measures of educational quality. It also sheds light on some of the most common problems underlying the design of teacher evaluation systems and barriers to successful implementation of teacher incentive programs. Several considerations regarding teacher accountability that derive from this literature are discussed at the end of this chapter.

INCENTIVE PROGRAMS: THEORY AND ASSUMPTIONS

Education is a labor-intensive endeavor, and in most countries around the world, teacher salaries account for 60 to 95 percent of educational spending.[12] In many countries there is mounting concern that salary schedules based solely on education and seniority weaken teachers' incentives to exert effort and improve student performance. In addition, it has been argued that across-the-board salary increases tend to increase costs with no commensurate improvement in student performance (Lavy, 2002). To tackle both of these problems simultaneously, the use of targeted incentives has often been proposed.

The rationale behind using incentives to promote changes in individual behavior was first explored in the economics and business literature. Incentive programs in general, and pay-for-performance programs in particular, are designed to solve the employer's problem of motivating high performance among its workers when individual effort and ability is not readily measured or observed (Asch, 2005). In more technical terms, incentives are useful when a principal and an agent (i.e., an individual performing work for an employer) have differing objectives in a context of asymmetric information. An attractive feature of these programs is that the employer need not dictate a particular procedure for attaining the outcome.

[12] In Mexico this number is estimated to be around 90 percent (Santibañez, Vernez, and Razquin, 2005).

Education activities pose a similar kind of principal-agent problem. School administrators, parents, and policymakers all want teachers to achieve certain outcomes (e.g., improve learning, develop well-adjusted adults, build citizenship). However, they cannot monitor teachers' daily activities nor do they know what exactly needs to be done to achieve these objectives in a way that can be effectively communicated to and implemented by all teachers. To resolve this problem in a way that improves educational outcomes, researchers and policymakers increasingly favor incentive programs in education. Advocates argue that teachers face weak incentives to improve performance because pay is determined by educational attainment and seniority, neither of which has been found to have significant positive effects on student achievement, after a certain threshold has been attained (Rivkin, Hanushek, and Kain, 1998; Hanushek, 1996).

There are certain assumptions that underlie incentive programs in general that do not completely hold in education. First, incentive program design assumes that there is a clear and known definition of performance (Klerman, 2005). This implies a common agreement about outputs. In most examples from the economics and business literature, the output commonly deals with simple tasks for which there is a readily available measure for performance. For example, papers in the business literature focused on incentives that raised productivity for workers installing car windshields (Lazear, 2000) and planting trees (Paarsch and Shearer, 2000). This is not the case in education. Teachers and schools have multiple and complex objectives (e.g., learning, student self-esteem, citizenship, development of core values, social skills. In addition, the objectives of teachers and their employers (principals, school boards, the government) are not necessarily aligned. For example, government agents might be interested in ensuring an equitable education for all its citizens, while teachers might want to keep difficult students out of the classroom (Vegas and Umansky, 2005).

Second, incentive programs assume that incentives will encourage workers to perform better, and that workers know the best way to achieve the intended outcomes. Research on accountability in education, however, suggests that educators might need help figuring out how to improve student achievement (Hamilton, 2005). This could be one reason why many accountability and merit-pay systems include provisions for technical assistance to schools that fail to meet targets. However, as Hamilton (2005) notes in her essay on lessons from performance measurement in education, even this technical assistance might not be enough to compensate for insufficient capacity problems (e.g., insufficient material and financial resources) or for the broader context in which some schools must operate. Even in the United States, where extensive research has been conducted on instructional practices, there is a general lack of knowledge about how to improve practice aside from a few well-documented findings (Hamilton, 2005).

Third, incentive programs assume that the benefit of obtaining a desired outcome outweighs the costs of measuring performance. This is less of a problem when simple tasks are involved, but when performance is multidimensional and involves a series of complex tasks (as is the case in education), measurement costs might outweigh the benefits of the incentives (Asch, 2005). This conundrum relates to the above-referenced premise of incentive programs — that is, that they presuppose an operational definition of performance (Klerman, 2005). The more precise and operational this definition, the easier it will be to design a system of incentives to reach it.

Because education is a multidimensional task, focusing on a single dimension of the educational equation (e.g., student test scores) could lead to problems such as curtailing creative thinking or teaching to the test (Hannaway, 1992; Holmstrom and Milgrom, 1991). There are ways to curtail these unintended responses, for example by combining the use of outcome measures (e.g., student tests) and practice measures (e.g., classroom observations, interviews, supervisor ratings). There has been some evidence to suggest that using repeated incentives improves the quality of the performance measure. Because there is certain to be some level of “noise” affecting each period’s evaluation, rewarding the worker for performance during a single period might expose him to considerable risk (Prendergast, 1999). In addition, it might result in the program rewarding workers whose performance improved due to a positive external change in working conditions rather than increased efforts in a given year. However, the costs and feasibility of using more extensive measures need to be compared with the likely benefits of adopting a more comprehensive system.

Before concluding this section, it is worth mentioning that some argue that explicit incentives (e.g., salary incentives) can reduce productivity by eliminating intrinsic desire. Most of the incentives literature in business and economics assumes that effort is costly for the worker. Other research (mainly in the psychology and sociology literatures) argues that in some cases the pride or sense of mission workers derive from their work makes carrying out their tasks an enjoyable activity. Teaching might just be such a job. Teachers often cite a desire to work with young people and contribute to society as important reasons to enter the teaching profession (see Guarino et al., 2004 for a review of this literature in the United States). However, some studies, e.g., Pendergrast 1999, suggest teachers have little intrinsic motivation. Regardless, there is little empirical evidence that incentives reduce intrinsic motivation. Therefore, it is fair to conclude that there is no solid evidence on whether explicit incentives in education contribute to creating a climate that hinders the intrinsic benefits of the job, although this is certainly an avenue of research worth exploring.

In sum, the literature reviewed in this section suggests that most of the theoretical underpinnings explored in the economics literature on incentive programs are not necessarily sustained in education. However, there are some cases in which teacher incentive programs have had positive effects on student achievement. These are discussed in the next section.

INCENTIVE PROGRAMS IN PRACTICE: EXPERIENCES FROM AROUND THE WORLD

Policy options to improve teaching around the world focus on improvement of three main areas: (1) teacher preparation and professional development; (2) recruitment and retention; and (3) teachers' work in the classroom (Vegas and Umansky, 2005). Many teacher incentive programs implemented in the past decade around the world focus on the third option (CM, to a certain extent, attempts to target all three areas), and most focus specifically on improving student test scores. Focusing on student test scores is an attractive choice because they provide observable measures of student learning, and there is evidence to suggest that they are correlated with other longer-term outcomes such as life-long earnings (Klerman, 2005). In this section we review some recent teacher incentive programs intended to improve student achievement as well as a few programs designed to meet other objectives. Most of these programs are small. Because few countries have implemented national-scale incentive programs (we are only aware of the Mexican and Chilean cases), the literature on such large-scale programs is limited.

Small-Scale Teacher Incentive Programs Designed to Improve Student Test Scores

Glewwe, Ilias, and Kremer (2003) examine the issue of teacher incentives in Kenya. They report on a randomized evaluation of a program that provided teachers with bonuses based on student performance. The program also aimed to reduce teacher absenteeism. In their data, teachers were absent from school 20 percent of the time and absent from their classrooms even more frequently. Teachers in 50 rural schools were randomly selected (out of 100) for the incentive program based on the average test score of students already enrolled in school at the start of the program. Each year the program provided prizes valued at up to 43 percent of typical monthly salary to teachers in grades 4 to 8 based on the performance of the school as a whole on the Kenyan government's districtwide exams. The authors found that during the life of the program, students in treatment schools were more likely to score higher, at least on some exams, than students in the comparison group. An examination of the channels through which this effect took place, however, provides little evidence of greater teacher effort aimed at preventing dropouts or increasing long-run learning: student dropout rates did not fall, teacher attendance did not improve, the amount of homework assigned did not increase, and pedagogy did not change. There is, however, evidence that teachers increased efforts to increase the number of pupils taking tests in the short run and to raise short-run test scores (by focusing instruction on the test). In addition, the authors found evidence that teachers adjusted to the program over time by offering more test preparation sessions (Glewwe, Ilias, and Kremer, 2003).

Lavy (2002) examined a small-scale program in Israel that provided incentives to teachers in 62 secondary schools. He used a regression-discontinuity approach to compare student outcomes in these schools with those of schools that just missed treatment because of eligibility rules. The program was implemented in 62 nonrandomly selected secondary schools in Israel during 1995. The program offered combined incentives to schools in the form of performance awards, part of which were distributed to teachers and school staff as merit pay and the rest used for the well-being (upgrading general work conditions) of teachers in the school. The total sum awarded was determined in advance (about $1.4 million) and was distributed among the top third of performers. Awards were given to teachers based on outcomes such as test scores and dropout rates. This type of incentive program is of the “rank-order” type, which awards prizes based on the rank order of the winners.

An interesting feature of the Lavy (2002) paper is that it compares an incentive intervention with what he calls a “resource” intervention, that is, where all schools selected for the program are given additional (and identical) resources to improve. The resource intervention that Lavy analyzes consisted of a three-year program that endowed 22 high schools with additional resources, mainly teaching time and in-service training. This resource program’s objective was to improve students’ performance, and the 22 participating schools (selected from 75 applicants) planned and developed the intervention individually and independently. A comparison group of schools not admitted into the resource program serves as the basis for identification of programmatic effects. Lavy evaluates the effect of this parallel program and compares its effectiveness and cost to the teachers’ incentives intervention.

Lavy’s results suggest that teachers’ monetary incentives had some effect in the first year of implementation (mainly in religious schools) and caused significant gains in many dimensions of students’ outcomes in the second year (in all schools). The program led to an increase in average test scores and a higher proportion of students who gained the high school matriculation certificate (especially among those from a disadvantaged background). It also appears to have contributed to a decrease in the dropout rate in the transition from middle to high school. The results regarding the resource program suggest that it also led to significant improvements in student performance. However, the comparison of the programs based on cost equivalency suggests that the teachers’ incentive intervention is more cost-effective (Lavy, 2002).

Dee and Keys (2004) analyze one incentive program implemented in the United States in Tennessee. The Career Ladder Evaluation System implemented in Tennessee to improve student achievement blended salary rewards with nonpecuniary benefits such as increased professional responsibilities (e.g., supervising beginning teachers, curriculum development). The program emphasized teachers’ professionalism, as opposed to simply providing some teachers with more money. Using results from the Tennessee STAR class-size experiment, which collected assessment data, the authors found that the Career Ladder Evaluation System program had only mixed success in targeting rewards to the more meritorious teachers. They concluded that assignment to a career-ladder teacher significantly increased mathematics scores by roughly three percentile points. However, most career-ladder teachers were not significantly more effective at promoting reading achievement. Furthermore, assignment to a teacher who had advanced further up the career ladder was not uniformly associated with significantly higher achievement (Dee and Keys, 2004). The authors acknowledge that one important caveat to their analysis stems from the fact that it uses an assessment system that was specifically designed to evaluate class size and not the effects of the incentive program.

Teacher Incentive Programs Designed to Improve Other Outcomes

Because most of the programs implemented recently deal only with effects on student scores, the impact of targeted teacher incentive programs on other outcomes such as teacher professionalism and professional development participation is less clear. There have been, however, some projects implemented in Mexico and elsewhere that use incentives to improve teacher attendance and teacher training and to attract and retain teachers, including those working in rural areas. We review several of these here.

Programa para Abatir el Rezago Educativo (PARE) was implemented in Mexico from 1992 to 1996 as part of a World Bank initiative to provide additional resources to very disadvantaged students, primarily those in rural areas and indigenous communities. One of the program's primary goals was to reduce teacher absenteeism in rural areas, which at the time was believed to be rampant. The instrument used to target this problem was teacher incentives in the form of monetary bonuses offered to teachers who could prove that they attended school regularly. Teachers were given a monetary incentive if they did not miss more than just a few days of school during the academic year. A novel feature of this program was that it would not be the teacher or the school principal reporting on the teachers' attendance record, but the parent associations. Most of these associations were created in response to another one of PARE's incentive plans, which offered parental associations bonuses of US$500 per year to encourage more involvement in school activities.

A qualitative evaluation of PARE, conducted in nine schools in two Mexican states, found that indeed teacher absenteeism was a problem. Teachers in the studied cases were in their classrooms for only 100 out of the 200 days in the school calendar. Furthermore, even when in school, teachers usually taught only two to three hours of the four and one-half hours in the school day. Evaluators found that the incentives effectively reduced teacher absenteeism only in schools that had "strong" school principals and parent associations or had low levels of teacher absenteeism *before* PARE was implemented (Ezpeleta and Weiss, 1996).

There is some evidence to suggest that programs that tie monetary incentives to acquisition of professional development or teacher certification have achieved some success. One such study is the evaluation of the National Board Certification Pilot Project in Iowa. This project offers teachers monetary incentives in exchange for obtaining advanced teacher certification. The evaluation found that teachers involved in the certification process with the monetary incentives engaged in more professional development activities than teachers not targeted by the incentives (Dethlefs et al., 2001).

In Mexico, Ornelas (2002), in his study of CM, concluded that one of the major successes of the program's monetary incentives was to emphasize professional development. "Hundreds of thousands of teachers now partake in professional development courses (offered by the federal government, through the national professional development program or PRONAP, and state authorities. After 1994 (with the implementation of CM), more supporting books and materials as well as textbooks have been printed along with books on educational policy and pedagogical theory; hundreds of courses are now offered via satellite to all teachers who wish to take them; the teacher centers (*Centros de Maestros*) were created each with a library containing more than six thousand volumes, along with videos and CDs. Most of the professional development going on right now takes place at these centers" (Ornelas, 2002, p. 20).

However, it is not difficult to make the case that professional development should not be a goal in and of itself, but should be an intermediate objective to improving student learning. While it is important for teachers to receive continuous training, it is more important that training results in improvements in teaching that have ultimate beneficial consequences in the classroom.

On the issue of how incentives affect teacher recruitment and retention, there is considerable U.S.-based evidence suggesting that increasing teacher salaries affects who chooses to enter and remain in teaching. In the United States numerous studies suggest a link between higher teacher salaries and higher retention (Podgursky et al., 2004; Stockard and Lehman, 2004; Kirby, Berends, and Naftel, 1999; Gritz and Theobald, 1996; Brewer, 1996). A few studies in the United States have also found that higher salaries are associated with higher quality entrants into the profession (Figlio, 2002; Loeb and Page, 2000). For example, using data from the Schools and Staffing Survey for 1987–88 and 1993–94, Figlio (2002) found that districts that raised their salaries relative to other districts in their county increased the possibility of hiring new teachers (both first-time and experienced transfer teachers) from more selective undergraduate institutions and with college majors in their teaching field.[13]

Incentives used to attract teachers to less desirable schools may not work as well. In Bolivia, local authorities offered teachers in rural areas a salary bonus to compensate them for the hardship of living in a rural community. Because of urbanization and demographic growth, some areas classified as rural were in fact borderline urban. An evaluation of this program found that the test scores and other educational outcomes of students of urban- and rural-classified teachers with the same background and characteristics were not significantly different. Furthermore, this evaluation concluded that the rural pay differential had not been successful in attracting and retaining more effective teachers to rural areas (Urquiola and Vegas, 2005).

If, by their mere presence, incentive programs result in salary increases for teachers and succeed in making the profession more attractive to more able or competent individuals, then it is possible that incentives might achieve positive objectives beyond the narrow focus on student test scores. It should be noted, however, that in the Mexican case, the labor market for teachers functions in a very different way than most private sector or even public sector labor markets. In Mexico, teacher education programs (as well as teaching positions [*plazas*] and wages) are centrally controlled by the federal government with strong involvement by the teachers’ union through negotiations regarding the teacher education curriculum. Therefore, it is possible that the labor market for public school teachers in Mexico is not able to respond (which might be the case in private schools, which are not subject to these regulations) to changes in salary. Put another way, because teaching positions in public schools are so rigidly controlled, it is not entirely clear that

[13] Figlio (2002) notes that these results held for districts that unilaterally raised salaries relative to those in the surrounding districts and might not generalize to a situation in which salaries are increased in all districts within a large geographical area.

higher salaries would automatically (or even in the medium term) result in a greater supply of more qualified individuals willing to enter teaching.

Group Incentives

Not all incentive programs are targeted to individual teachers. Some programs use group incentives, such as targeting the school. Schoolwide initiatives could overcome perverse consequences resulting from individual incentives (Dee and Keys, 2004). Some authors argue that merit pay for teachers and other such programs distort the incentives for a variety of teacher behaviors such as cooperation and collective effort, and foster an unproductive work environment (Murnane and Cohen, 1986). This might be particularly true of rank-order incentive programs that rank teachers against one another.

In response to these criticisms, some countries have moved away from individual incentives and have implemented group interventions (e.g., Lavy´s 2002 study of Israel). Another example is Chile's *Programa de Mejoramiento de la Calidad de las Escuelas Básicas de Sectores Pobres* or *Programa de las 900 Escuelas* (P900). This program targets the lowest performing public schools in the country. It allocates resources to schools with a mean score below a certain threshold.

The program began in 1990 and provides technical and material support to teachers, directors, and students in four main areas: (1) improving teacher quality through periodic workshops; (2) providing student workshops designed to help students at-risk to raise their grades and enhance their self-esteem, social skills, and creativity; (3) preparing and distributing textbooks, creating classroom libraries and educational material; and (4) improving school management through supervised work groups with teacher representatives, directors, and community authorities that design the institutional education program best fitted to the school's objectives and the community (Tokman, 2002). In her evaluation of the program Tokman (2002) concluded that it had in fact narrowed achievement gaps between low performing and average or above-average performing schools (Tokman, 2002). Other authors, however, found that these positive effects were mainly due to a mean-reversion effect. Once mean-reversion was accounted for, the program's effect remained positive, but its magnitude decreased considerably (Chay, McEwan, and Urquiola, 2005).

Another Chilean program, the *Sistema Nacional de Evaluación del Desempeñode los Establecimientos Educacionales* (SNED) or National System of School Performance Assessment, offers monetary bonuses to schools that show high student-achievement marks. This program was implemented in 1990 and preliminary evidence shows a cumulative positive effect on student achievement for schools which had a relatively high probability of winning the award (Mizala and Romaguera, 2003).

ADVERSE AND UNINTENDED CONSEQUENCES OF TEACHER INCENTIVE PROGRAMS

Some of the studies that we reviewed did link incentives to improved teaching practices. Most of the research on teacher incentives, however, does not explicitly investigate the mechanisms through which incentives work to make teachers change their behavior and ultimately affect student outcomes. Some studies that have tried to explore these mechanisms have found that incentive programs had adverse or unintended responses. For example, Jacob and Levitt (2003) found that improvement in student test scores observed in Chicago public schools after the introduction of teacher accountability reforms was due to cheating. Other studies have found that exclusion of low achieving students (Cullen and Reback, 2002; Figlio and Getzler, 2002), focusing only on subjects that are tested (Hamilton, Stecher, and Klein, 2002) teaching to the test (Koretz, 2002) or even increasing students' caloric intake on the day of the test (Figlio and Winicki, 2005) were largely behind improvements in student outcomes. These findings are sobering and illustrate the challenges associated with targeted incentives.[14]

Group incentives can also have adverse consequences as they have a potential for "free-riding." Research has shown that if an employee's share of the reward is small relative to the difficulty of the work, and if the effort of all team members is difficult to observe by the employer, an individual on the team will have an incentive to shirk his or her work while benefiting from the teams' work or free-riding on the effort of others. This phenomenon weakens the power of group incentives (Asch, 2005, Prendergast, 1999; Asch and Karoly, 1993).

[14] Although there are no rigorous investigations to prove it, some authors argue that these gaming behaviors also occur in Mexico to improve student test scores on the CM tests (García Manzano, 2004; Ornelas, 2002).

CONCLUSIONS

The evidence regarding the impact of programs that offer teachers monetary incentives to improve student achievement is mixed although it does suggest small improvements in the short run. Most of the programs reviewed in this chapter did show modest improvements in student test scores, but these were often short-lived and were not always the product of increased teacher effort or improved teaching practice. In some cases the positive results are derived from interventions that feature both individual and group (or school-level) incentives. Moreover, most of these incentive programs have only been in existence for a few years and their longer-term impact is unknown. Incentives used for other purposes, such as increasing teacher attendance, have demonstrated mixed results. The evidence on the positive effects of higher salaries on teachers engaging in professional developments (when incentives are offered to engage in such activities) and on teacher recruitment and retention seems to be more solid.

It is worth noting that none of the programs we discuss here have penalties for poor performance. In addition, it is not clear whether these programs achieve equity in schools. For example, teachers in better schools could be disproportionately receiving a larger share of rewards than teachers in worse schools if the rewards are based solely on student achievement. If the incentive system reduces morale among teachers who do not receive rewards, this could leave students in failing schools at an even greater disadvantage.

Last, all the programs reviewed here used monetary incentives to encourage teachers, but nonmonetary incentives are also a possibility. In education, teachers might respond to incentives in the form of improved working conditions, in-kind contributions (e.g., educational materials), job stability, pensions, and the like (Vegas and Umansky, 2005). Other common examples of nonmonetary incentives are promotion-based (where promotions to higher grades in a career schedule are based on performance assessed over several time periods) and seniority-based incentives (when employers offer a reward later in an employee’s career contingent on current levels of effort) (Asch, 2005). Both of these can be effective in improving worker productivity, but they are of limited use in educational systems.[15]

[15] Obviously, the first kind of incentives only works in an organization with a vertical hierarchy (e.g., workers promoted to managers). The second approach suffers from the problem of workers not wanting to retire when they

are senior employees because they are being paid more than their productivity (Lazear, 1979; 1983). In addition, this approach is in some ways reflected by current salary schedules that are based on seniority. Research has found that more senior teachers are not necessarily more competent than junior teachers who have been teaching for at least two to three years (Hanushek, 1997). But senior teachers receive higher salaries than novices do on all salary schedules.

3. DESCRIPTION OF CARRERA MAGISTERIAL

This chapter provides a description of Carrera Magisterial (CM). First, it describes the historical context that made CM possible. Second, it describes CM by outlining its main design features and operation guidelines. Here, we discuss CM's program timeline, eligibility rules, and the range of the salary bonuses. This section also presents national data on teacher participation as well as changes in assessment scores over time.

HISTORICAL CONTEXT

To understand the climate for reform that brought about the implementation of CM in 1993, it is important to recognize the role played by the teachers' union (SNTE) and the particular political climate of that time.[16] During the 1980s, Mexican teachers saw their living and working conditions deteriorate considerably. The decline in their real wages led to teacher strikes and dissident movements in 1989, resulting in new leadership of SNTE (Ornelas, 2002; Ávila and Martínez, 1990). At the same time, educational decentralization was a key piece of President Salinas' 1988 political agenda (Ornelas, 2002). CM's design and implementation owe a great deal to both of these competing agendas: increasing wages (on the part of SNTE) and modernizing education (on the part of the Ministry of Education).

After a long and difficult negotiation, the *Acuerdo Nacional para la Modernización de la Educación Básica* (ANMEB) was signed into law in 1992 by the federal government, SNTE leaders, and state authorities. The signing of the agreement led to CM's implementation. According to Ornelas (2002), SEP wanted to implement a true merit-pay system to reward the best teachers. The teachers' union wanted to implement a new kind of horizontal salary schedule that would reward all teachers equally. SEP wanted to include only classroom teachers in the program, while SNTE wanted the program to cover its entire

[16] Since its founding in 1943, the national teachers' union, or *Sindicato Nacional de Trabajadores de la Educación* (SNTE), has played a significant role in Mexican education and politics. By law, there was only one teachers' union in Mexico for decades. Basic education teachers contribute around 1 percent of their monthly wages in union dues. In 1990, SNTE had about one million members. Its large membership and the amount of the contributions made by teachers gave SNTE great political and economic power.

membership. When it was signed, the agreement had elements of both proposals, but most of SNTE's demands prevailed (Ornelas, 2002).

MAIN PROGRAM FEATURES

CM is jointly managed by the Ministry of Education (SEP) and SNTE. The program's governing body is the *Comisión Nacional* SEP-SNTE, which consists of eight members of the Ministry of Education and seven members of the teachers' union. The SEP-SNTE Commission is a fundamental part of the program. Centrally designed program guidelines, assessment instruments and factors, and membership criteria all resulted from negotiations between SEP and SNTE. In the field, the program is operated by state governing bodies, or *Comisiones Paritarias* (which are also made up of SEP and SNTE members at the state level). These commissions are in charge of determining final incorporations into and promotions in the program, publishing program results, and negotiating conflicts.[17]

ELIGIBILITY

To be eligible for CM, teachers must teach in a public primary or secondary school in Mexico. They must have a minimum seniority requirement (two to six years) that varies by educational attainment. Therefore, most new teachers are not eligible. In addition, both teachers and principals must hold specific kinds of contracts (*plazas*) with SEP.[18] Teachers with two teaching positions can participate in the program under both positions.[19] Secondary teachers on contracts with fewer than 19 hours per week are not eligible to participate.

[17] Guidelines for state commissions are included in the document *Guía Técnica para Determinar el Número de Plazas a Incorporar o Promover en el Programa Nacional de Carrera Magisterial*.

[18] Teachers must have a contract that is "*Código 10 (alta definitiva)*", which is equivalent to a tenured position, or a "*Código 95 sin titular (interinato ilimitado)*" which is equivalent to a nontenured, but unlimited temporary contract. These are teachers that are using somebody else's *plaza*. Many teachers take a leave from teaching but do not renounce their tenured positions. These are then made available to SEP for allocation among new teachers. Teachers who receive these positions, though, do not officially own them, even though they might work under that contract for their entire teaching careers.

[19] If a teacher participates with two positions, he only needs to take the teacher test once, provided that both of the positions are in the same primary cycle or secondary subject.

To be eligible for incorporation into the program (meeting the criteria for the program's first level, Level A) or promotion (meeting the criteria for subsequent higher bonuses at Levels B to E), teachers must obtain a score in each of the six assessment factors. Failure to complete the assessments on all six factors results in a total point score of zero. In some cases, this requirement can be waived. For example, if a state does not offer state professional development courses, teachers cannot be assessed on this factor. However, if a teacher's students fail to take the student test, he will automatically obtain a final score equal to zero. Teachers will also get a score equal to zero if they fail to take the teacher test (which also means that the Dirección General de Evaluación (DGE) will not test their students).

FACTORS AND SCORING

CM rewards teachers via salary incentives, on the basis of education, ability, experience, training, and student knowledge. These variables are measured by the program's factors. Table 3.1 describes the factors (or measures of performance) used by CM as well as the points attached to each. The total score in CM is obtained by adding all the factor point scores up to a maximum of 100 points.

Years of service (seniority) and the highest degree earned are converted into a point score based on an established formula set by the program (see the national program guidelines, annexes 8 and 9, *Comisión Nacional* SEP-SNTE, 1998).[20] Professional development is evaluated by a test that all teachers who completed state and national courses must take. Professional performance is evaluated by a survey completed by the principal and other teachers to rate various dimension of a teacher's practice. Student achievement is measured by the teacher's students' results on a standardized test. Teacher knowledge and skills are evaluated by a standardized test on the relevant subject matter,

[20] Note that not all education degrees are accepted. CM only considers degrees granted by teacher colleges (*Escuelas Normales*), upper teacher colleges (*Escuelas Normales Superiores)*, universities and institutes of higher education with official (government) recognition and included in the National Registry of Institutions Belonging to the National Education System (*Comisión Nacional* SEP-SNTE, 1998). In practice this implies that a teacher with a degree from a non-Mexican institution will not be able to receive any points in this factor.

teaching methodologies, and the Mexican education system and law.[21] Teacher, student, and professional development tests vary only by grade and subject, not by CM level. All of these instruments have been developed specifically for CM.

Table 3.1
Teacher Promotion Factors in Carrera Magisterial (CM): Original (1993) and Current (1998) Guidelines*

Factor	1993 Max. Points Awarded	1998 Max. Points Awarded
1. Seniority	10	10
2. Educational Credentials	15	15
3. Professional Preparation (measured by a teacher test)	25	28
4. Training and Other Professional Development Courses	15	17
5. Professional Performance (supervisor/peer review)	35	10
6. Student Achievement (measured by a student test)	7**	20

Source: *Comisión Nacional* SEP-SNTE, 1998; 1993.
* The original 1993 CM guidelines were changed in 1998 to reflect new program priorities and adapt to changes in the education system.
** These points were included in the 35 assigned to Professional Performance.

At the end of the evaluation cycle, raw scores are transformed into CM point scores (those in Table 3.1) using a linear transformation intended to allow for fair comparisons among teachers. To compute the student achievement CM score and the teacher test CM score, the Evaluation Directorate DGE, (which

[21] In 1998 the program underwent a major reform after a comprehensive study of teachers' perceptions of the program carried out in 1996 found inconsistencies and loopholes in the program guidelines (*Coordinación Nacional de Carrera Magisterial*, 2001). Prior to 1998, student academic achievement and professional performance were collapsed into a single factor that awarded teachers up to 35 points. The number of points a teacher received was largely based on a peer-review instrument, and not on the student test score (which contributed only up to seven points of the 35 possible). The 1996 review found that the majority of teachers had received the maximum number of points in the professional performance factor. Because the mechanism of peer review was not thought to be reliable, in 1998 the professional performance factor was separated into a student achievement factor that yielded up to 20 points, and a peer-review component that contributed up to 10 points. The remaining five points were distributed among the professional preparation factor (i.e., the teacher's test score) which was awarded three additional points, and the professional development factor with two additional points (see Table 3.1).

does the scoring for these two factors) divides teachers in groups according to state, grade (or subject), type of school (e.g., regular, technical, indigenous, distance) and strata to denote the development of the region where the school is located.[22] Teachers are ranked within each group from the highest raw score to the lowest. The teacher or teachers with the highest score obtain 20 points for student achievement and 28 points for teacher knowledge and the rest get a lower number of points relative to their raw scores.[23]

TIMELINE OF ASSESSMENTS

CM's assessments and other activities are conducted in a one-year period known in program terminology as an *Etapa,* or program year. The program begins in October when teachers sign up and enroll. They are asked to provide proof of their highest degree earned and years of service. National professional development tests are administered in November, teacher tests in March, and student tests in June. Teachers receive their total CM scores as well as incorporation or promotion outcomes in September (this concludes the *Etapa*). Table 3.2 describes the timeline for CM's major assessment activities.

[22] These strata denote marginal rural areas, marginal urban areas, low development rural areas, low development urban areas, and medium development urban areas.

[23] There are some exceptions to this rule. First, teachers whose students are not tested (first and second grade in primary, and teachers of single-teacher, indigenous, and distance schools, as well as secondary teachers in some subjects), get a student test score that is scaled from their teacher test score. Second, secondary teachers who teach different subjects (within the same grade or in different grades) will only get one student evaluation and it will be for the subject and grade that accounts for the most hours of work. In addition, teachers will only receive the teacher test for the subject with the highest number of assigned hours. They get to choose the subject in the case of a tie in hours worked.

Table 3.2
Timeline of CM Activities

Month	Activity
October	Teachers enroll (sign up) in the program. Teachers who enrolled in previous years receive a prefilled form that includes their highest degree earned and seniority. Registration for PRONAP (National Program for Teacher Professional Development) courses. (End of October) First professional performance (peer review) evaluation period.
October–May	State courses are delivered. Within this period states decide what courses are given and when. (Most courses last about a week.)
November	National professional development courses are delivered by PRONAP, which also administers professional development tests.
January	Second professional performance (peer and supervisor review) evaluation period.
February-March	PRONAP sends course score reports to teachers.
March-April	Teacher tests (professional preparation) are administered and scored by the Directorate of Evaluation (DGE). Results are sent to CM.
May	PRONAP sends course scores to CM.
May-June	States send course scores to CM.
June	Student tests are administered by DGE and scores are sent to CM. (Mid-June) Third professional performance (peer review) evaluation period. (End of June) Peer review score reports are sent by schools to CM.
September-October	CM sends teachers detailed final score report. Evaluation results as well as promotions and incorporations are published. Period for appeals begins.

Note: This program timeline can vary somewhat from year to year.

CM LEVELS, INCORPORATIONS, AND PROMOTIONS

There are five levels of promotion in Carrera Magisterial, denoted A, B, C, D, and E, each with a successively larger salary bonus. They must be pursued sequentially, beginning with Level A. A move to Level A is referred to as "incorporation." Moves into Levels B and beyond are referred to as "promotions." We will use that terminology throughout this report. Note that the tests do not get more difficult at each level. All teachers take the same test (based on their grade and subject) regardless of their CM level. Therefore, even though the salary bonus that a teacher in Level E receives is almost 10 times the bonus received by a teacher in Level A, the tests that led to that teacher receiving a promotion to that level are the same tests that allowed another teacher to enter the program's lowest level. Moreover, professional development courses are also identical for all levels of teachers and vary only according to grade and/or subject. Using the same assessments at all levels calls into question the assumption that those teaching at the highest levels of CM are the best teachers.

Once incorporated or promoted to a given level, individuals cannot be demoted and their wage bonus remains constant throughout their careers. However, after their incorporation they must wait for a specified number of years before they can attempt promotion to higher levels.[24] During their waiting period, teachers can take the assessments every year and use their highest scores when they become eligible for a promotion.

Promotion opportunities are determined by each participant's final point score, calculated by the national CM office. Final point scores are distributed via an electronic spreadsheet to each of Mexico's 32 state-level offices where the final selection is conducted. In the early years of the program, states were given much leeway in determining the cutoff score above which teachers were promoted, and these apparently varied by state, and from year to year within states. Starting in *Etapa* 9 (the 1999–2000 school year), incorporations into the program became based on a well-publicized national cutoff score (Ortiz Jiménez, 2003).[25] Currently, no individual scoring below 70 points is eligible for incorporation into Level A. Promotions into Levels B to E are not subject to this cutoff (i.e., there is no minimum score).

Actual incorporations and promotions depend on states' allotted budgets for CM incentives. Each year, as part of its education budget, SEP makes transfers to the states to pay for CM's salary bonuses. States can choose what

[24] As an incentive for teachers in low-development areas, they only need to remain in the program for two years before becoming eligible for promotion into the next level. Teachers in areas not labeled as low development must remain in the program for three to four years (depending on the level) before they can be promoted.

[25] Up until the program's eighth year, it appears that there were few rigid rules governing each state's approach to allocating the awards. Each state was allocated a yearly budget that constrained the number of promotions that could be awarded. Once the number of promotions was determined, it appears that most states relied heavily on the final point score to allocate rewards. However, it is not clear whether states adhered to a deterministic cutoff (i.e., all teachers scoring above a given point value would receive a promotion).

proportion of the budget to allocate to promotions and what proportion to incorporations. On average, states devote 75 percent of the resources to promotions and 25 percent to incorporations.[26] They also choose how much to devote to each *vertiente*, that is, teachers, school administrators, and support staff (teacher aides). During *Etapas* 8 to 11, states dedicated about 85 percent of their CM budget to teachers and the rest to administrators and school support. This is roughly equivalent to the number of teachers in relation to administrators and support professionals in each state.

SALARY BONUSES

By using salary bonuses as an incentive mechanism, CM seeks to "help improve the quality of education in Mexico through recognizing and supporting the work of the best teachers …" and "reinforcing interest in teacher professional development and continuous improvement" (*Comisión Nacional* SEP-SNTE, 1998). Monetary incentives offered by the program are considerable. A teacher's Level A bonus is more than 20 percent over the base salary (determined from a traditional salary schedule using seniority and education). Teachers in Level E receive over 200 percent over the base (see Table 3.3).

These bonuses are in effect over the life of the teacher's tenure, and teachers do not lose their Level A benefits even if they are never evaluated again. The bonuses are large when compared to teacher incentive plans in other countries, particularly given the fact that they are not simply one-time payments.[27]

[26] Currently one of the most important aspects of the yearly salary negotiation between SEP and SNTE are the amounts that will go to fund CM incentives. This becomes a center piece of the negotiation because often times the increases to base salary are relatively low and are designed mainly to compensate for inflation. More information about federal and state CM budgets is available at SEP's webpage www.sep.gob.mx, under "Carrera Magisterial."

[27] For example, Glewwe, Ilias, and Kremer (2003) note that incentive plans in the U.S. typically offer one-time payments of 10 to 40 percent of the monthly wage. Israel has offered payments of 60 to 300 percent of the monthly wage, but these were still only one-time payments (Lavy, 2002).

Table 3.3
CM Salary Bonuses by Level (in thousands)

	TOTAL GROSS (MX$)	Base Gross (MX$)	CM Bonus (MX$)	Bonus/ Base(%)
Primary				
Level A	7,570	5,971	1,599	27
Level B	9,554	5,971	3,583	60
Level C	12,199	5,971	6,228	104
Level D	15,088	5,971	9,117	153
Level E	18,787	5,971	12,816	215
Secondary				
Level A	7,271	5,747	1,524	27
Level B	9,232	5,747	3,485	61
Level C	11,704	5,747	5,957	104
Level D	14,558	5,747	8,811	153
Level E	18,209	5,747	12,462	217

Source: http://www.sep.gob.mx/wb2/sep/sep_tabuladores (positions E0280 and E0362).
Note A: Secondary bonuses are for secondary teachers teaching at least 19 hours per week.
Note B: Amounts listed are monthly salary levels.

The program features we have discussed above paint an ambiguous picture of CM, a result perhaps of contradictory visions by SEP and SNTE. The fact that salary bonuses are retained over a teacher's entire career leads us to question whether CM truly is a system of teacher incentives or whether it is merely a new salary schedule where the nonincorporated group of teachers constitutes the lowest level. Although the program's objectives is to reward better teachers, the program's design practically ensures that an average teacher, one with the basic teaching degree (*Normal Licenciatura*) who remains in service for a sufficiently long time, attends professional development, and obtains an average score (equal to the national mean of 10 points) in both the teacher and student test, will eventually gain incorporation into the program.

PROGRAM PARTICIPATION AND TRENDS IN ASSESSMENT SCORES

Every year, CM conducts statistical analyses of program participation and assessment scores.[28] This section presents a general picture of teacher participation in the program as well as of some trends on teacher scores. In

[28] See http://www.sep.gob.mx/wb2/sep/sep_618_coordinacion_naciona.

this section we present mainly national numbers. Disaggregated tables by development area and teacher characteristics can be found in Appendix A.

Program Participation

Table 3.4 depicts teacher participation figures from 1991 to 2002. During this time, 51 to 60 percent of all basic education teachers in the country participated at some point in CM.[29] Note that a participating teacher is any teacher who signed up to become evaluated by CM at the beginning of the school year.

Not all teachers who enroll and participate in CM are actually evaluated. CM counts as evaluated teachers only those who complete all factor assessments. This proportion has suffered a considerable decline since 1991. Table 3.4 shows that in 2002 only 70 percent of participating teachers were actually evaluated.

[29] Even though the official statistics often do not make an explicit note of this, when the Ministry of Education reports “number of teachers” it is actually reporting *plazas*, *not* individual teachers. A *plaza* is defined as a teaching position and it can be temporary or permanent. A permanent *plaza* (or *plaza de base*) is a teaching position with tenure. Once a teacher holds a *plaza de base*, he has a legal right to remain a teacher in the Mexican school system for life. In fact, teachers own their *plazas* and can inherit them or even sell them (in some extreme cases). Many secondary teachers are on hourly contracts, and may hold 5, 10, 20, or up to 42-hour *plazas,* which entitle them to teach at particular secondary school for the number of hours specified in their contract. If a teacher holds a morning contract in one school and an evening contract in the same (or a different) school, he is officially counted as two teachers. Informal reports from the Mexican Ministry of Education suggest the proportion of teachers in the Mexican public school system with double contracts in primary schools is somewhere between 60 to 75 percent. This figure might even be higher for secondary school teachers who work by hours and thus usually have a few *plazas* for a different number of hours each.

Table 3.4
Teacher Participation (Enrollment), Assessment, and Incorporation into CM (*Etapas* 1-11): All *Vertientes*

	Total Teachers	Participating Teachers (Enrolled)	Participating Teachers/ Total Teachers (%)	Evaluated Teachers	Evaluated Teachers/ Participating Teachers (%)	Newly Incorporated Teachers	Newly Incorporated Teachers/ Participating Teachers (%)	Total Incorporated Teachers (cum.)*	Total Teachers Incorporated/ Total (%)
1991-93 (*Etapas* 1-2)	838,750					389,816		389,816	46
1993-94 (*Etapa* 3)	863,042	442,334	51	406,214	92	13,700	3	403,516	47
1994-95 (*Etapa* 4)	894,076	522,973	58	469,069	90	58,391	11	461,907	52
1995-96 (*Etapa* 5)	894,076	553,625	62	486,483	88	40,066	7	501,973	56
1996-97 (*Etapa* 6)	946,505	569,171	60	499,260	88	32,033	6	534,006	56
1997-98 (*Etapa* 7)	959,013	583,259	61	515,166	88	26,518	5	560,524	58
1998-99 (*Etapa* 8)	982,925	610,542	62	527,080	86	10,468	2	570,992	58
1999-2000 (*Etapa* 9)	995,486	618,441	62	486,945	79	9,210	1	580,202	58
2000-01 (*Etapa* 10)	1,013,647	594,616	59	450,257	76	8,964	2	589,166	58
2001-02 (*Etapa* 11)	1,028,524	615,721	60	431,775	70	5,772	1	594,938	58

Source: Coordinacion Nacional CM, Estadistica Basica *Etapas* 3-11. Available at: http://www.sep.gob.mx/wb2/sep/sep_618_coordinacion_naciona. *Etapas* 1-2 numbers come from Ortiz (2003). CM numbers include all *Vertientes* (i.e., teachers, principals, and aides). Source for total teachers (column 1) is SEP, *Estadística Básica* 1990-2000, 2000-2003. Note that the total teachers include public and private schools because that is how SEP reports them. However, since public schools comprise about 90 percent of all basic education schools in the country, CM participation rates are likely to be only slightly underestimated. All numbers refer to teacher positions.

* Many of these teachers subsequently moved on to higher levels (B-E). The cumulative incorporation numbers, therefore, should not be interpreted as reflecting the number of teachers in Level A at any given *Etapa* of the program.

Only a small percentage of teachers achieve Level A each year. In 2002, CM incorporated about 1 percent of total participating teachers into its first level, Level A. However, as of 2002, 58 percent of total basic education teachers in Mexico had been incorporated into CM's Level A. Note that not all of these teachers are currently at Level A, as many of them have been promoted into higher levels (Levels B–E) following their first incorporation.

The incorporation numbers in Table 3.4 demonstrate that two-thirds of teachers (389,816 out of 594,938 teachers) were incorporated during the first two *Etapas* of the program. During the first two years, CM's evaluation system was not yet fully functional. Therefore, these incorporations were not based on a formal assessment. Even though CM is designed to reward teachers based on merit, in practice, CM resulted in an across-the-board salary increase for most of the teachers currently in the program. Moreover, after *Etapa* 7, the proportion of incorporated teachers relative to total teachers in the country remained stable. This could suggest that the budgetary restrictions are in fact working to restrict program membership growth. Another explanation could be that the policy to base points on more stringent assessments resulted in more competition.

Table 3.4 shows incorporation rates only (column 8), but average promotion and incorporation rates have been stable since *Etapa* 8 at around 6 percent every year (2 percent incorporations, 4 to 5 percent promotions). Close to half the promotions are from Level A to B. About 40 percent are promotions from Level B to C. Promotion and incorporation rates vary slightly between states and economic development areas. Economically developed areas (i.e., urban areas of medium development) have slightly higher incorporation and promotion rates, while rural and marginal areas have the lowest rates (see Appendix A).

As previously discussed, any enrolled teachers who fail to obtain scores in all six factors are not counted as "evaluated." We conducted an analysis to determine if there are certain factors that teachers often fail to complete. The results from this analysis are reported in Table 3.5.

Table 3.5
Exploration of Teachers with Total Point Scores Equal to Zero

	Primary		Secondary	
	No. Positions	**%**	**No. Positions**	**%**
Total positions	1,788,627		402,468	
Positions with total point scores equal to zero	903,833	50	212,236	53
Positions with total point scores = 0 & PD scores equal to zero	803,132	89	205,507	97
Positions with total point scores = 0 & PD scores not equal to zero	100,701	11	6,729	3
Of those with total scores = 0				
Everything positive except PD	352,448	39	67,491	32
Everything positive except student or teacher test	5,992	1	31,060	15
PD, student and teacher tests = 0	336,324	37	89,023	42
All factors = 0	53,113	6	15,164	7
Other cases	155,956	17	9,498	4
Total		100		100

Note: PD scores include scores on either the national or the state professional development test taken by the teacher. All numbers refer to teacher positions.

Table 3.5 demonstrates that between *Etapas* 8 and 12, close to 90 percent of teachers with total point scores equal to zero failed to obtain a positive score on either the national or state professional development courses. Most of these teachers did not obtain a positive score because they did not take the course. In fact, of all teachers getting a zero total point score, 40 percent got positive scores on all factors except for professional development. The remaining cases mostly consist of teachers who neither took professional development courses nor completed an assessment of other factors. The correlation between failing to take the teacher test and failing to take the student test is very high, reflecting CM's policy of not testing students of teachers who fail to take the teacher test.

Table 3.6 shows the number of teachers currently at each of CM's five levels. These figures are from the dataset used in this study and thus include only teachers participating in CM at some point between *Etapas* 8 and 12. Therefore, teachers who gained an incorporation or promotion to the program prior to *Etapa* 8 and/or who during *Etapas* 8 to 12 did not sign up to become evaluated, are not part of these statistics. As Table 3.6 demonstrates, the majority of CM's teachers are in Levels A and B, and most of those are in the first level. Few teachers are in the higher levels of D and E. Note that during the early years CM used two levels, BC and CC, which have now been discontinued.[30]

Table 3.6
Number of Teachers by CM Level (*Etapas* 8 -12)*

Level	Primary	%	Secondary	%
A	96,909	60	29,129	73
B	48,881	30	8,692	22
BC	4	0	0	0
C	13,171	8	1,349	3
CC	2	0	0	0
D	3,640	2	566	1
E	207	0	16	0
Total	162,814	100	39,752	100

* These numbers are cumulative as of *Etapa* 12, for teachers showing up in the CM database between *Etapas* 8–12. All numbers refer to teacher positions.

Trends in Assessment Scores

Total teacher point scores for both primary and secondary teachers increased slightly between *Etapas* 8 and 12. Disaggregated figures by teacher or regional characteristics do not show large differences in scores by gender, but they do show some differences by seniority, education, and region (see Appendix A for full results). With respect to teacher test scores, at the primary level there is no clear pattern for teachers with more or fewer years of seniority. Secondary teachers with more seniority do appear to have slightly higher teacher test scores than those with fewer years in the system. However, there are large improvements in teacher test scores for both primary and secondary teachers with more education compared to those with less.

With respect to student test scores, primary teachers' student scores increase until a teacher reaches about 10 years of seniority. There is no clear pattern of relationship between teachers with greater seniority levels and student achievement. Neither is there a relationship between primary teachers’ level of education and student achievement. At the secondary level,

[30] These were intermediary levels between B and C and C and D that were given to teachers with special circumstances (e.g., those who had graduated from the *Universidad Pedagógica Nacional*).

students of teachers with more seniority scored slightly higher on average than students of less senior teachers. As was the case at the primary level, there is not a relationship between level of education and student achievement at the secondary level.

In terms of region, primary teachers who work in urban areas classified as having medium levels of development obtain almost identical student test scores on average as teachers who work in urban areas of low or marginal development.[31] However, teachers who work in urban areas of medium development do obtain slightly higher teacher test scores than teachers in low and marginal development areas, although the difference is very small.

Table 3.7 shows global point score averages by CM level (note that we only report average scores after *Etapa* 9, that is, after the minimum cutoff of 70 points for incorporation began to be enforced for the majority of the participating teachers).

Table 3.7
Mean Total Point Score by Level of CM (*Etapas* 9-12)*

	Primary		Secondary	
Level	Global Point Score*	Score w/o Seniority or Education	Global Point Score*	Score w/o Seniority or Education
A	71	56	63	47
B	80	64	85	67
C	82	65	86	68
D	83	65	87	68
E	83	64	89	67

*This is the mean global point score used for incorporation or promotion into that level. Seniority and education points are taken from the year the teacher position is observed.

When scores in factors that increase each year automatically (seniority) or have little variance (highest degree earned) are removed, average scores rise until Level B and then remain fairly constant. This stability might be due to the fact that the same tests are used as teachers progress through CM (although this would not explain the difference between Levels A and B, only between Levels B and beyond).

[31] This variable had a lot of missing data at the secondary level and thus was not used for the analysis.

CONCLUSIONS

Mexico was one of the first countries to explicitly tie the salary of all public school teachers to teacher performance. CM is a national, centralized program, operated jointly by the Ministry of Education (SEP) and the teachers' union (SNTE). Perhaps due to the tension between these two actors and their differing visions, CM's nature is ambiguous. It lies somewhere between a merit-pay system and a new salary schedule.

CM's design gives teachers the opportunity to prove their knowledge and teaching abilities on a variety of indicators, including seniority, education, and professional development, as well as teacher and student tests and peer-review ratings. Teachers who enter the first level (Level A) and are incorporated receive a bonus of more than 20 percent over the base salary (determined from a traditional salary schedule using seniority and education), while teachers in the program's highest level (Level E) receive over 200 percent over the base.

The majority of teachers in Mexico participated at some point in CM (i.e., signed up to become evaluated) and are currently in one of the program's five levels. Participation in CM is not universal because not all teachers are eligible. Increasingly, teachers enroll in the program but fail to take professional development courses (and the tests associated with this factor). This results in an annulment of the evaluation and the teacher receives a total point score of zero. A little over half the teachers in *Etapas* 8 to 12 faced this situation. It might be worth examining whether or not this situation is due to a lack of professional development courses available to all teachers in all regions of the country (which would constitute an access barrier to incorporation and promotion), or whether it simply represents individual teacher's choices.

Because one of the program's objectives is to reward the better teachers, we would expect teachers in the highest levels of CM to be among the "best." However, there are no major differences in CM total point scores between teachers once they reach Level B. Ascension to Levels C and beyond is mostly determined by point scores obtained in factors that either increase automatically (e.g., seniority), or have little variance (e.g., education). Furthermore, it is important to underscore that the teacher and student tests do not become more difficult as teachers progress in CM. Teachers teaching the same grade and subject receive the same test (and the same professional development courses and tests), regardless of whether they are in Level A or Level E. These programmatic components could lead one to doubt that teachers in the upper echelons constitute the most competent teachers.

4. ANALYSIS OF CARRERA MAGISTERIAL'S STUDENT AND TEACHER TESTS

One of the assumptions of Carrera Magisterial (CM), presented in graphical form in the introduction to this study, is that student achievement and teacher test results constitute or represent adequate measures of what students and teachers know or should know. Despite the considerable amount of resources, time, and effort involved in the testing of millions of students and thousands of teachers, however, surprisingly little is known about the quality of the CM tests and their content, and the procedures followed in their development. It is important that testing programs provide evidence attesting to the technical characteristics and overall quality of tests, especially if the test scores are used to inform high-stakes decisions, such as promotions or bonuses (AERA, APA, and NCME, 1999). The purpose of this chapter is to examine the technical quality of the teacher and student tests used as part of the CM program. We do this by addressing the following research questions:

1. Does the content of the teacher tests match the test specifications provided to teachers?
2. Which subjects/topics are most and least emphasized by the primary grade tests? Which topics are most and least emphasized by the secondary grade tests?
3. What are the levels of cognitive demand elicited by the teacher and student tests at the secondary level?[32]
4. What do item statistics indicate about the functioning of the items in the teacher and student tests?
5. What do test statistics indicate about the internal consistency (i.e., reliability) of the teacher and student tests?
6. How do CM's test development and administration procedures fare with respect to international standards of practice for educational testing and measurement?

The first research question addresses concerns that the content guidelines teachers receive for the teacher test might not accurately reflect the tests' actual content. The second question relates to the content covered

[32] Due to the reduced sample of items from every subject in the primary tests, this analysis could only be performed on the secondary tests.

by the tests. It is important to characterize the content assessed on a test as it can send signals to examinees about which skills are more important (Education Trust, 1999). Research has shown that more attention is directed toward topics and subjects that are tested than those that are excluded (Shepard and Dougherty, 1991). We examine two aspects of test content, namely the subjects and specific topics included in the tests.

The third question explores the levels of cognitive demand in the teacher and student tests used in CM. Underlying current curricular reforms in reading and mathematics (and other subjects under development) is a movement away from memorization and toward conceptual understanding, which seems to imply at least an equivalent level of cognitive demand for teachers. Moreover, CM is part of a broader effort by SEP to move away from the idea of the "teacher as technician" to that of the "teacher as a professional." This intuitively implies that teachers should have specialized knowledge and the ability to perform tasks that involve a higher level of cognitive demand (i.e., assessment, evaluation, feedback, adaptation of pedagogical practices, and so forth).

The fourth and fifth questions investigate the statistical properties and general technical quality of the tests. Item discrimination indicates the extent to which an item distinguishes the higher-performing from lower-performing teachers (or students). Low discriminating items contribute little to the overall effectiveness of the test and may indicate problems in wording, content, or coding. It is also important to understand whether items that are designed to measure similar themes or constructs in fact elicit similar responses from examinees. The more consistently examinees respond to items measuring similar constructs, the higher the reliability of the test. Tests with higher internal consistency are desirable because they provide more stable estimates of examinees' proficiency than tests with lower internal consistency.

The sixth and final question analyzes whether CM's testing program conforms to internationally accepted standards of practice for educational testing. This is important given the size of the testing program that CM operates, and the prominent role the teacher and student tests have in incorporation and promotion outcomes for teachers, which ultimately have large effects on salaries.

TEACHER AND STUDENT TEST ANALYSIS

This section provides a summary of the methods and findings on the CM teacher and student test analysis.

Data and Methods

The analysis presented here is based on five sources of data: the tests themselves, the test specifications (study guides) provided to teachers for the teacher test, the content guidelines or standards (*planes de estudio*) specified in the national curriculum, summary item and test statistics provided by the ministry’s evaluation office (DGE) for the teacher and student CM tests used in 2003–04, and the item-level dataset containing the student responses to the items in the primary and secondary tests for the same year.

We analyze teacher tests used for primary grades 1 to 6, and student tests used for primary grades 3 to 6. At the secondary level, we focus on Spanish, mathematics, history, geography, and civics and ethics for the three secondary grades (7 to 9). To guide the analysis of test content vis-à-vis the curriculum standards, a framework was developed for each subject summarizing the major topics in the standards which was then analyzed by a group of expert and nonexpert raters (with knowledge of the Mexican Education System).[33]

To explore the levels of cognitive demand elicited by the secondary-level teacher and student tests, we used frameworks that specified several cognitive categories and asked the experts to make judgments on the cognitive demand of the items. The cognitive categories were adopted from several sources, including test frameworks (e.g., National Assessment of Educational Progress, NAEP), as well as coding categories used in studies of alignment (e.g., Webb, 1997; Achieve, Inc., 2004). While the cognitive categories necessarily vary by subject, interest centered in examining emphasis in lower- and higher-level skills, irrespective of the particular subject assessed by an item. We used Bloom’s taxonomy (Bloom et al., 1956) to obtain a hierarchy among the

[33] We convened expert consultants for each subject, all fluent in both Spanish and English, and with undergraduate degrees in the relevant subject. Half the experts also had experience teaching the relevant subject. Experts were asked to provide feedback on the comprehensiveness of the framework relative to the curriculum standards and on the feasibility of the framework for coding the test items. The proportion of agreement among expert and novice raters is used to explore the reliability of the coded responses. The same group of expert and novice raters was used to examine whether the content of the teacher tests matched test specifications provided to teachers.

cognitive categories. Bloom’s taxonomy specifies six categories of cognitive processes, arranged in order from simple (e.g., literal recall) to complex (e.g., evaluation).

Last, we examined the following statistical properties of the tests items:[34]

- Difficulty index: the proportion of examinees that answer an item correctly in a given administration of the test.
- Discrimination index: the difference in the proportions of total correct responses between the highest-scoring and lowest-scoring examinees.
- Item-total point-biserial correlation: the correlation between examinee responses to a particular item and the total scores obtained on all other items.
- Cronbach’s alpha (Cronbach and Shavelson, 2004): the ratio between item variance and the variance of the total scores adjusted by test size. It equals the average correlation among test items.

Since CM does not intend to produce reliable indicators of individual student achievement, the key statistic in the case of the student tests is the reliability of the classroom averages used in teacher promotion and incentive decisions (Cronbach, Linn, Brennan, and Haertel, 1997; Brennan, 1995). These statistics, unfortunately, are not estimated by CM and thus were not available to us. Item-level data provided by the DGE were used to obtain reliability estimates at the classroom level that take into account the item matrix sampling design.

Findings

Table 4.1 summarizes the results of the teacher test analysis. In terms of the structure and reliability of the teacher test, the content section of the teacher tests generally exhibited adequate or nearly adequate levels of reliability (except on the civics and ethics test). The sections on teaching methodologies and law have lower estimates of internal reliability. This is partly due to the few items used to assess knowledge of educational law. While there is variation across subjects, the average item difficulty is generally

[34] Teacher tests were constructed to assess three distinct domains: subject-specific knowledge, pedagogical skills, and regulatory aspects of educational law. Item statistics were analyzed separately by domain.

adequate (on average, between 55 and 60 percent of the examinees answered the items correctly). Content sections of the tests are generally more difficult, with a mean index of 0.56 compared to pedagogical knowledge: 0.62 and educational law: 0.61.

The teacher tests have broad coverage of most curriculum areas. In all cases the tests are well aligned with test specifications. However, our evidence suggests that most of the items in the content section of the teacher test assessed low-level cognitive skills. While these results should be taken cautiously, they suggest that explicit attention and further exploration by SEP and researchers in general are needed regarding the levels of cognitive demand required by the teacher tests.

Table 4.1
Summary of Results: Teacher Test Analysis

	Overall Finding	Primary Grades (grades 1-6)	Secondary Grades				
			Spanish	Math	History	Geography	Civics/Ethics
Content matches teacher test specifications	Yes						
Overall content coverage (subjects/topics more and less emphasized)	Broad	Not very broad (emphasis is on Spanish and math; Mexican and world history are least emphasized)	Broad (least emphasis on library and reference skills, speaking and listening)	Broad (least emphasis on prealgebra and basic algebra)	Broad	Broad	Results not reliable*
Levels of cognitive demand	Low	N/A**	Low (most items assess literal recall)	Low (most items require demonstration of procedures)	Results not reliable*	Results not reliable*	Results not reliable*
Functioning of test items and test difficulty	Adequate (few items had poor discrimination); adequate difficulty	Adequate functioning and difficulty	Adequate functioning and low difficulty	Adequate functioning and difficulty	Adequate functioning and low difficulty	Adequate functioning and difficulty	Poor discrimination (one-quarter items flagged); low difficulty
Internal consistency (reliability of the test)	Mixed	Adequate for content, low for pedagogy and educational law	Adequate for content, lower for pedagogy and educational law				Low

* Results are deemed not reliable when agreement between expert and novice coders was low. This disagreement suggests that the cognitive framework used may need to be improved or that training exercises might need to be revised before conducting future analyses.

** There were too few content-specific items in the primary tests to undertake this analysis.

Before discussing the student test analysis, we contrast our teacher assessment results with the literature on international assessment practices. The CM subject-matter section of the teacher test assesses teachers' knowledge of the curriculum they teach. Although there are theories and empirical evidence to support the link between basic skills in mathematics, reading and writing, and student achievement (NRC, 2001), many of the tests used in other countries to evaluate teachers are much broader. Porter, Youngs, and Odden (2001) in their review of teacher assessments and their uses, classify the types of assessments into those that assess: basic literacy and numeric skills, subject matter knowledge, pedagogical knowledge or knowledge of how students learn, teaching skills (clinical practice), and gains in student learning.

As they further describe, assessments of basic skills or competencies are frequently used to make decisions of initial licensure or of entrance into teacher education programs. Licensure decisions are often made on the basis of tests assessing subject-matter knowledge and pedagogical knowledge (similar to the teacher tests in CM). Once teachers obtain a license to teach and are engaged in classroom work, however, assessments of continuous practice follow a more in-depth and subjective approach (such as principal ratings or ratings made by trained assessors using a previously developed instrument) that typically involve completing class and instruction profiles, interviews and classroom observations. Although none of these assessments are without problems (e.g., the high costs of in-depth examinations, compressed ratings, the limited range of what can be assessed with a paper and pencil test), they might constitute a more logical approach to evaluating teachers as they progress from teacher education to initial licensure and continuing practice.[35]

[35] To counteract criticisms of traditional assessments, some countries, like the United States, have developed more in-depth assessments of teacher practice such as the Praxis I, Praxis II, and Praxis III test series developed by the Educational Testing Service. Praxis I is used for entrance into teacher education programs or initial licensure. It measures basic skills in reading, writing, and mathematics. Praxis II is used to determine initial licensure. Both include assessments of core content knowledge and sections assessing more in-depth content as well as teaching and pedagogical content knowledge. The core tests follow a multiple-choice format, while the in-depth tests feature open questions. Praxis III is used for continuing license decisions and to evaluate teaching skills. This assessment is based on 19 areas of practice that were deemed important by teachers, researchers, state officials and

In addition, we have not seen anything in the literature that would support testing teachers on their knowledge of the legal, administrative, and other aspects of the educational system as does the legal section of the CM teacher test.

Next, we discuss the tests used to evaluate student achievement. Table 4.2 summarizes the results of the analysis of the student tests.

In the case of the student tests, these have broad coverage of the curriculum, with the primary tests and the civics and ethics in secondary being the exception. As was the case with the teacher test, the student test demands relatively low levels of cognitive skills, although these tend to increase in the higher grades. This result was also found by Schmelkes (2001) in her study of CM using a sample of teachers in marginal development areas.

Last, the matrix sampling design used in the student tests is sound for the purposes of maximizing content coverage and reliability of classroom aggregates. In terms of the item and test statistics, the results indicate that a large proportion of items have low discrimination values. In terms of difficulty, the secondary history, mathematics, and geography tests are considerably more difficult than the other tests (index close to 40 percent) — perhaps reflecting the larger proportion of items measuring higher levels of cognitive demand. The civics test, on the other hand, is probably easier than intended. Finally, the alpha coefficients indicate that many of the tests have lower levels of internal consistency (i.e., reliability) than would be desirable, especially in the case of the secondary tests.

feedback from a national advisory committee (Dwyer, 1998). In Praxis III teachers complete a class profile and instruction profile and are observed and interviewed by trained evaluators. Because of the high costs associated with Praxis III, a simplified version known as Pathwise is used for formative purposes, although not to make licensure decisions (Porter, Youngs, and Odden, 2001).

Table 4.2
Summary of Results: Student Test Analysis

	Overall Findings	**Primary Grades (grades 3-6)**	**Secondary Grades**				
			Spanish	**Math**	**History**	**Geography**	**Civics/Ethnics**
Overall content coverage (subjects/topics more and less emphasized)	Broad	Not very broad (emphasis on Spanish and math)	Not very broad (emphasis on reading and writing; least emphasized: library skills, reflections on literature, spoken language and other forms of communication)	Broad	Broad	Broad	Not very broad (least emphasized are topics in ethics and society)
Levels of cognitive demand	Low to moderate	N/A*	Mostly low (most items require low skill levels (half of the items assess literal recall), although some items require analysis)	Low to moderate (mostly low in grades 1 and 2, higher in 3)	Low (most items require literal recall of historical data)	Balance between low, moderate, and high skills	Low (most items require identification or description only)
Functioning of test items and test difficulty	Low (most items had poor discrimination values); adequate difficulty	Adequate discrimination and difficulty	Adequate discrimination and difficulty	Adequate discrimination; more difficult	Adequate discrimination; more difficult	Adequate discrimination; more difficult	Poor (one-quarter items had poor discrimination); low difficulty
Internal consistency (reliability of the test	Mixed (refers to classroom-level averages)	Adequate (classroom-level averages)	Low (classroom-level averages)				

* There were too few content-specific items in the primary tests to undertake this analysis.

** In all cases, classroom-level aggregates had higher levels of internal consistency than those exhibited by the student-level results. However, since the program aggregates student-level results to the classroom level, we only report internal consistency indicators at this level.

ANALYSIS OF CM’S TESTING PROGRAM

This section describes the results of our analysis of CM’s testing program, including its test development system, test administration, and organizational structure. The goal of this analysis is to determine whether CM’s testing program operates according to minimum benchmarks of quality accepted by the measurement community worldwide.

Data and Methods

Our review was based on three sources of data. The first was CM’s program guidelines and documents from the Ministry of Education’s Directorate of Evaluation (DGE) describing the process of development and administration of the teacher and student tests. Importantly, the level of detail in these documents varied greatly, and often no documentation was available at all. To fill in gaps in the information needed to properly evaluate CM’s testing procedures, we interviewed key DGE personnel in Mexico City during May 2005, constituting our second source of information. The interviews provided crucial insight into the process followed in the development of the teacher and student tests when information was not available in writing and when it was insufficient. The third source of data consisted of published technical work in the field of educational assessment and measurement.

To determine whether CM’s testing program operates according to minimum benchmarks of quality accepted by the measurement community worldwide, we employed three sets of standards. First, the joint Standards for Educational and Psychological Testing are widely used internationally to guide the development and evaluation of standardized tests (AERA, APA, and NCME, 1999). In addition, we considered the evaluation framework proposed by the National Research Council (NRC, 2001) as it specifically addresses issues of quality in teacher licensure tests. Finally, we referenced the Quality Standards for Educational Evaluation Instruments (ECIEE) (Martinez-Rizo et al., 2000) developed in Mexico by an international committee of experts. The ECIEE standards build primarily upon the AERA-APA-NCME standards and the Program Evaluation Standards (Joint Committee, 1994), complementing these with other important work in the area of educational testing and evaluation in an effort to produce a set of standards relevant to the Mexican context.

As the authors of the joint standards stress, not every standard is relevant to all tests. Instead, the importance of each standard or series of standards for a particular test depends "on the context and purpose of the test development or use" (AERA, APA and NCME, 1999). Consequently, the first step in our analyses was to review the AERA, NRC, and ECIEE standards and identify the most relevant standards for evaluating the teacher and student tests used in CM. We grouped these as they relate to distinct phases and components of the testing program:

- *Test contents and test specification* concerns the process of planning the test and delineating its general goals and characteristics, as well as determining the specific contents, constructs, and cognitive dimensions to be measured.
- *Test construction* includes the process of item writing, piloting, and revision, and the construction of the test based on the content specifications and statistical considerations and criteria.
- *Reliability* refers to the statistical properties of the tests, specifically internal structure and consistency, and item discrimination. *Validity* is not a property of the test itself but refers to the degree to which the intended uses of a test are supported by empirical evidence and educational theory.
- *Test administration* refers to the procedures for test delivery, including selection of examinees or groups of examinees to be tested, timing, accommodations, and mechanisms for fraud prevention.
- *Fraud prevention and test security* includes the mechanisms and procedures in place to ensure the integrity of the test materials and information during and after the development and administration process.
- *Information and communication* involves the mechanisms in place to provide information to interested parties prior to administration of the test. It also includes regulations involving the release and distribution of test results to examinees and the public.
- *Organizational structure* refers to formal components usually in place in a large-scale testing program, such as content and

technical advisory committees, which contribute to and oversee the quality of the tests produced.

We analyzed the documentation available for the teacher and student tests and the information obtained in the interviews, using the joint standards as our primary reference (AERA, APA, and NCME, 1999), the NRC framework as support for reviewing the teacher tests, and the ECIEE standards (Martinez-Rizo et al., 2000) when they provide insights relevant to the Mexican context.

Findings

Table 4.3 presents a summary of findings of the analysis of CM’s testing program by each of the categories of teacher testing program standards described above. Detailed AERA, APA and NCME, ECIEE, and NRC standards referred to in this table can be found in Appendix B.

Our evaluation of the teacher and student tests used in CM in light of international standards of practice produced mixed results. The teacher tests, while still in need of improvement, appear to be more carefully developed and monitored than the student tests, which are of comparatively lower quality.

In test construction, the tests are given different levels of attention. The student tests are constructed without explicit content tables or specifications and the items are written and informally revised within SEP. Consequently the tests are assembled without any concrete evidence of item quality. Tests and items are never piloted.

Table 4.3
Analysis of CM's Testing Program: Summary of Findings

	Student Tests				**Teacher Tests**			
Standard Group	**Available Documentation***	**Improvement Needed****			**Available Documentation**	**Improvement Needed**		
		Minor	**Moderate**	**Major**		**Minor**	**Moderate**	**Major**
Test Contents and Specification *(AERA 3.2, 3.3, 3.5, 3.7, 3.11)*	Limited			X	Partial		X	
Test Construction *(AERA 3.6, 3.9, 3.11, 3.17, 7.3; ECIEE 4.2, 4.3, 14.2)*	Limited			X	Partial		X	
Reliability *(AERA 2.1, 2.2, 2.7,2.19 3.9,4.1,4.2,5.12,13.19; ECIEE 5.2, 8.2)*	Partial			X	Limited			X
Validity *(AERA 1.1, 1.3, 4.3, 4.17, 11.2,15.7; ECIEE 6.1)*	Limited			X	Limited			X
Test Administration *(AERA 3.19, 5.1, 5.2; ECIEE 10.4, 10.5)*	Adequate	X			Adequate	X		
Test Security *(AERA 5.6; ECIEE 9.10, 9.11m, 10.6)*	Partial		X		Partial		X	
Fraud Detection *(AERA 8.11; ECIEE 11.6, 12.8)*	Partial		X		N/A			
Information *(AERA 6.1, 8.1, 11.5; ECIEE 9.1, 9.2)*	Partial		X		Adequate	X		
Communication *(AERA 11.5, 5.10, 11.15, 11.18; ECIEE 14.3, 14.10)*	Limited			X	Partial	X		
Organizational Structure *(AERA 1.7, ECIEE 1.1)*	Partial			X	Partial		X	

*Definition of documentation availability categories: *Limited*: Minimal or nonexistent documentation. *Partial*: Some documentation available but not sufficient to be deemed adequate. *Adequate*: Adequate documentation is available.

**Definition of improvement categories: *Minor*: Requires very few changes or adaptations to procedures/test characteristics, or organization of existing documentation. *Moderate*: Requires moderate to high adjustment of problematic procedures/test characteristics, or completion of procedures and documentation. *Major*: Requires an extensive revision of problematic practices/test characteristics and in some cases the development of nonexistent procedures, policies, and documentation.

By comparison, items in the teacher test are created according to content tables and specifications. While the teacher tests also consist largely of new unpiloted items which are reviewed by committee after administration, an increasing number of previously piloted items are available in a computerized bank developed to aid in the construction of the tests; equating of teacher tests across years is now underway.

Nevertheless, we found significant limitations in the process to ensure technical quality of the teacher tests. For example, formal procedures are not in place, nor are any statistical analyses conducted to review items for different kinds of biases (e.g., gender, cultural, regional). A study conducted in Mexico using a national sample of teachers working in rural and marginal development areas found that the CM tests are biased toward more urban environments (Schmelkes, 2001). Furthermore, budget constraints have forced SEP to stop conducting item-writing workshops in recent years so that, as with student tests, items are largely developed and reviewed in-house. The problems of quality control in test construction result in student and teacher tests that contain many poorly functioning items.

There is no technical manual describing the statistical and psychometric characteristics of either the teacher or student test, and information about the reliability of the tests is not available in any one document. To look at the reliability of the student test, SEP uses student-level estimates of reliability, but it is reliability at the classroom level that should be estimated, as this is the level used for CM assessment purposes. Even though classroom-level reliability estimates should be higher (on average) than student-level estimates, the previous section indicates that classroom-level reliability remains inadequate in some cases. Finally, no information is available about the properties of the scale transformation applied to the test scores to produce CM points.

The evidence on validity is minimal for both the teacher and student tests. Validity analyses in CM are limited to judgments of face validity made by item writers, DGE personnel (often these are one and the same), or the SEP-SNTE Commission. No studies are available on the internal factorial structure of the tests (the findings from the previous section suggest many tests may suffer from problems of multidimensionality), the relationship of test scores to external criteria, or on any other evidence of the validity of the test scores for the purposes of CM.

In the case of student tests, the scores are used for purposes that differ from those that the tests were designed for, without any empirical evidence to support the validity of these uses. For example, aggregate school and teacher level results are reported to states for diagnostic purposes based on the performance of the students taking the CM tests in a particular year. This use of the CM student tests is not formally tied to teacher incentives through CM but can have considerable impact on schools and teachers (particularly if the sample in the states that year is not representative of the teacher population).

Moreover, states increasingly use rankings of CM scores to gauge the state of their educational system in comparison with other states. The self-selected nature of the sample, both of CM participants and of teachers tested each year, should preclude such comparisons. Furthermore, some states use these aggregates to analyze time trends in the achievement of their students, even though these scores are reported in a z-scale based on the yearly mean and standard deviations and thus do not allow for estimation of trends across time (in other words, tests are not equated across time). As a result, even if we were to assume perfectly reliable tests it would be hard to ascertain whether the level of achievement at a school or state went up or down except perhaps for those with the most dramatic fluctuations from one year to another. If we add to this that the sample of students might not be representative of the school's or state's student population, we can conclude that statements of improvements or declines in student achievement at the school or state level based on CM test results are inadequate. While a brief note of caution is included in the annual reports suggesting the trends are relative rather than absolute (DGE, 2004), this contrasts with the very practice of reporting such trends. It appears to be cause for concern that the student tests are being used for purposes entirely different from those for which they were designed without documentation supporting the appropriateness of such practices.[36]

[36] The practice of reporting results from CM student tests at the school and teacher levels (via the SICRAE web service) is questionable. The system intends to provide diagnostic information to teachers, schools, and states about student and teacher status and progress using aggregate scores from CM's student achievement tests. The evidence presented in this study seems to raise doubt about the appropriateness of such practices. In any case, the validity

Administration of the CM student tests is a formidable enterprise that involves millions of students every year. DGE has developed and administers up to 45 different student tests, some with multiple forms. Relatively comprehensive guidelines and procedures for standardization and fraud prevention during administration of the tests have been developed and are adequately documented. However, security guidelines, when present, are not as robust as would be desirable in the periods before and after administration, as has been evidenced by recent security breaches. Guidelines about confidentiality and communication of information of test results and information to teachers, schools, state authorities, researchers, and the public are also lacking for both tests.

There is a general lack of adequate documentation about the tests and testing procedures — that is, test development procedures, item writing and revision, reliability and psychometric properties, test security, and information dissemination. When available, documentation is often fragmented and not formally compiled or published. One exception is the detailed guidelines available for administration of the student tests. However, the absence of technical manuals or other organized sets of guidelines and methodologies is a troublesome finding for a testing program of the size and scope of CM.

CM’s testing program seems to lack the organizational infrastructure desirable of a program of its size as well. For example, content and technical advisory committees are not set up to help improve and monitor the quality of the instruments. The Ministry’s office of curricular contents (DGMME) no longer participates in the specification of content for the student tests. Finally, item writing committees have not been used for the student tests and have been discontinued for budgetary reasons for the teacher tests.

Finally, it is important to note that in the absence of historical documentation, it is not clear that we can conclude that the features of the 2003–04 student test are representative of tests used in prior (or subsequent) years. The test development process discussed throughout this document could lead one to suspect that important fluctuations could occur from year to year in the psychometric characteristics of the student and teacher tests (e.g.,

of each type of use of the test scores should be thoroughly analyzed and documented (AERA, APA, and NCME, 1999).

item statistics, reliability, dimensionality). In addition to meeting minimum levels of measurement precision required for tests to be operational, an important recommendation for increasing the quality of the CM tests would be to monitor and maintain adequate documentation of relevant test statistics across years.

CONCLUSIONS

Administration of the CM testing program is a formidable enterprise that involves millions of students every year. The Ministry of Education through its evaluation office has developed the necessary human and technical infrastructure to administer dozens of CM teacher tests and over 45 student tests, some with multiple forms across the country. The fact that they can successfully administer these tests every year with a limited budget is commendable.

However, to improve the quality of the instruments, there are a few areas that deserve special attention. This analysis highlighted some areas of concern with the tests themselves, their technical properties, their levels of cognitive demand, and the administration procedures (note that we only analyzed tests for grades 3 to 6 at the primary level, and tests of Spanish, mathematics, history, geography, and civics and ethics at the secondary level).

Using international standards of assessment practice as a reference, the analysis of the teacher and student tests produced mixed results. Although there is room for improvement, the teacher test appears to be more carefully developed and monitored, the student tests appear to suffer from less attention and care in their development, which in turn results in tests of comparatively lower quality. Moreover, SEP does not have sufficient empirical evidence on the technical validity of both tests for test development. Analyses of validity of the CM tests appear to be limited to judgments made by the same people that write the items or the SEP-SNTE Commission.

The student tests are developed without explicit tables or specifications that relate to their content. The items are written and reviewed informally and consequently the tests are not constructed with solid evidence regarding the quality of the items. There is no piloting of tests or items. Nor are the appropriate reliability coefficients estimated at the classroom level. The reliability coefficients suggest that, particularly with the tests used at the secondary level, the tests show low levels of internal consistency—in some cases lower than what would be desirable. Moreover, the student test elicits mostly low levels of cognitive demand, although item difficulty does improve in the higher grades.

Improvements could also be made to the teacher tests. There are currently no formal procedures for item review to reduce potential biases (e.g., gender, cultural or regional). Moreover, because of budgetary restrictions, item-writing workshops have been suspended. Most items are now written and reviewed internally by SEP personnel. The evidence on cognitive demand suggests that most of the items in the content section of the teacher test assess low-level cognitive skills. In the case of both the teacher and the student tests, there is no available documentation regarding the technical properties of the formula used to transform raw scores into CM point scores.

In addition, we note it is challenging to vary the level of difficulty of teacher tests based mainly on subject knowledge for more senior teachers or those in higher levels of CM. A test based on competencies (e.g., analytic, evaluation, and critical thinking skills, and the ability to adapt one's teaching to different types of students and environments), as well as on subject-matter knowledge, could be better adapted to various levels of CM and better suited to identify highly competent teachers.

Last, CM's assessment system seems to lack the necessary infrastructure and resources for a program of its size. For example, the system lacks an advisory technical or content board that could supervise and help improve the quality of the tests. Overall, there is a lack of systematic and organized documentation on the tests and the testing procedures for development, administration, scoring, report, and use. When it is available, the information is often fragmented or is not formally published. The results also suggest that CM student test results are being used for purposes entirely different than those the tests were designed to serve and that this is done without any empirical evidence to justify these alternate uses.

5. ANALYSIS OF PEER-REVIEW INSTRUMENT AND PROCEDURES

Subjective evaluations of performance are generally believed to provide a better way to evaluate jobs that are complex and multidimensional such as teaching (Hamilton, 2005; Asch, 2005) than traditional assessments such as multiple-choice standardized tests of teachers' knowledge. In education, for example, carefully developed indicators that reflect complex dimensions of classroom practice are usually more powerful predictors of student achievement than traditional measures of teacher qualifications, experience, or knowledge (Schacter and Thum, 2004).[37]

Carrera Magisterial (CM) includes peer review as one of the six factors used to evaluate teacher performance. Peer review carries a maximum score of up to 10 points. This factor is evaluated using a rating instrument that assesses various dimensions of teaching. The evaluation is conducted by a committee composed of the principal, the union representative, and the teacher's peers (other teachers in the same school). For simplicity, we refer to this factor as the peer-review factor.

This chapter analyzes the instruments and procedures used for peer review by CM. We begin this section with a description of CM's peer review instrument and evaluation procedures. The instruments and procedures are assessed in light of relevant international research on the subject of supervisor and peer evaluations in education with a focus on six key elements: (1) definition of teaching standards; (2) acceptable documentation and levels of performance; (3) instrument reliability; (4) rating scale; (5) rating process; and (6) criteria for adding additional indicators.

[37] Peer-review processes in education usually focus on helping teachers to develop necessary competencies or address weaknesses. The process is usually led by consultants selected from a pool of distinguished and experienced teachers in the school or district. These teachers are usually freed from teaching loads partially or entirely to devote their full attention to serve as peer consultants for a number of years; they receive monetary compensation over base salary, training, assistance, and access to key state/district resources to help in their evaluator and/or mentoring role (Kumrov and Dahlen, 2002; Kelly, 1998). Peer consultants confer with teachers and can also conduct classroom observations; they provide mentoring, assistance, information, and guidance, and report on outcomes of the intervention in terms of teacher performance. Peer reviewers can also provide a summative evaluation or recommendation that can eventually lead to probation status, dismissal, or nonrenewal of contract.

CM'S PEER-REVIEW INSTRUMENT AND EVALUATION PROCEDURES

CM broadly defines professional performance as "the daily activities that teachers perform as part of their job" involving "classroom activities … essential for student growth" (*Comisión Nacional* SEP-SNTE, 2000). CM promotes the participation of teachers as evaluators of their peers because "they are witnesses of their daily activities" (*Comisión Nacional* SEP-SNTE, 2000). To evaluate professional performance, each school's technical council assumes the role of evaluation committee (*órgano evaluador*). The technical council is composed of all teachers in the school (in secondary schools it is composed of all full-time teachers), a representative of the teacher's union, and the school principal. The committee holds monthly meetings during the year to collect all the documentation and other elements needed to conduct the evaluation. The formal evaluation takes place at three points during the school year (usually in October, January, and June).

To evaluate professional performance, CM uses an instrument that measures 15 indicators.[38] These indicators correspond to four dimensions of teacher performance: (1) planning and execution of the teaching process; (2) planning and execution of the learning process; (3) participation in school activities; and (4) participation in the school-community interaction. The guidelines establish that although these dimensions do not exhaust all factors involved in a concept as complex as professional performance, it provides a "general estimate" of the extent to which teachers achieve desired educational goals (*Comisión Nacional* SEP-SNTE, 1998). Furthermore, the evaluating committee of each school can add up to five additional indicators to the evaluations of teachers. Additional indicators need to satisfy three criteria to be included in the evaluation: (1) reflect an important educational problem; (2) involve all staff participating in CM; and (3) be observable or measurable. During the three evaluation meetings, teachers present a written self-evaluation of their performance along with corresponding documentation or evidence to the evaluation body. After the committee deliberates (teachers under evaluation are not present during the deliberations), the committee assigns a rating to each indicator on a scale that can take three values: 1, 3, or 5. The final rating awarded to the teacher must be reached by consensus.

[38] A copy of the instrument (in Spanish) can be found in http://www.sep.gob.mx/wb2/sep/sep_cncm_instrumentos_desemp_profesional.

The first three indicators (related to instructional planning) are evaluated during the first evaluation meeting; the remaining indicators are evaluated during the second and third meetings. The second evaluation meeting does not produce ratings that formally enter into the final evaluation of performance; instead, it is intended to provide formative feedback to teachers so they can improve in areas that require attention. The third evaluation meeting is used to arrive at the final rating of the teacher's performance.[39]

ANALYSIS OF PEER-REVIEW INSTRUMENT AND THE EVALUATION PROCEDURES

Table 5.1 summarizes the findings of the analysis of the peer-review instrument and procedures.

Correspondence Between Review Process and Teaching Standards

Any evaluation of teachers' performance must be based on some accepted definition of teaching quality. Just as standardized tests provide users with accepted definitions of what should be known or what skills should be demonstrated, subjective peer evaluation must also be based on accepted standards for teaching. Ideally, teaching standards should be available with detailed definitions and examples to provide guidance to teachers about the behaviors and practices considered desirable (i.e., a model of what good teachers look like) for career advancement (Hamilton, 2005; Darling-Hammond, 2001).

[39] This rating (on a three-point scale of 1, 3, or 5) is then translated into CM points using the following formula: Professional performance score = (sum of scores in all indicators*2)/(number of indicators).

Table 5.1
Analysis of Peer-Review Instrument and Procedures: Summary of Findings

		Improvement Needed**		
Standard Group	**Documentation Available***	**Minor**	**Significant**	**Major**
Correspondence between review guidelines and existing teaching standards	Limited			X
Guidelines for documentation and evaluation of performance	Limited		X	
Criteria for additional indicators	Partial	X		
Rating scale	Partial	X		
Reliability of the instrument	No		X	
Validity of the instrument	No		X	
Peer review process	Partial			X

* Definition of documentation availability categories: *Limited*: Minimal or nonexistent documentation. *Partial*: Some documentation available but not sufficient to be deemed adequate. *Adequate*: Adequate documentation is available

** Definition of improvement categories: *Minor*: Requires very few changes or adaptations to procedures/test characteristics, or organization of existing documentation. *Significant*: Requires moderate to high adjustment of problematic procedures/test characteristics, or completion of procedures and documentation. *Major*: Requires an extensive revision of problematic practices/test characteristics and in some cases the development of nonexistent procedures, policies, and documentation.

By defining specific behaviors and benchmarks, teaching standards help teachers meet their stated goals, and facilitate the construction of indicators or measures of performance to evaluate these behaviors. CM’s peer-review evaluation, however, does not provide specific guidance to teachers and reviewers with respect to desired behaviors and evaluation criteria. This hinders the ability of CM’s peer review system to provide accurate measures of teacher quality and might help explain the heavily skewed ratings observed with this factor.

Research on instructional practices and teacher evaluation systems provide guidance on identifying and developing indicators of teacher quality (see, for example, Schacter and Thum, 2004; *Ministerio de Educación*, 2005). In Mexico, the national curriculum for elementary teachers (SEP, 1997) and secondary teachers (SEP, 1999) includes a description of “desirable characteristics of new teachers: profile at graduation." This profile constitutes a form of teaching standards related to five areas: intellectual abilities, content knowledge, didactic competencies, professional identity and ethics, and sensitivity to the school context. Although these standards are not operationalized into specific indicators, or complemented with levels of performance or a description of desirable behaviors, they constitute the framework of reference for evaluation of teachers in Mexico, as they are the clearest official statement of the desired characteristics of a good teacher. Moreover, this framework is much more comprehensive than the one used by the current CM guidelines for peer review. However, there does not appear to be a clear correspondence between these national curricular standards for teacher training and the guidelines for teacher evaluation followed in CM.

Guidelines for Documentation and Evaluation of Performance

One issue with subjective evaluations is that the accuracy of the assessment cannot be fully verified (Baker, 1992, cited in Asch, 2005). One way to minimize measurement error is to attach the evaluation to explicit standards of practice. For example, the Toledo Peer Assistance Plan in the United States (one of the pioneering peer-review programs in the country) addresses potential sources of unreliability by using evaluation criteria that are explicit and concrete, a small number of well-trained evaluators, and an evaluation process that promotes common assessment criteria and frequent observation and consultation with the teacher and the review panel.

Current CM peer review guidelines state that teachers must document their performance and activities through their own testimonies and any available documents. However, they do not include information about the desired teacher practices or behaviors each indicator is designed to measure, the rationale for using these indicators, or the criteria used to evaluate performance. The instrument and guidelines also lack specificity with respect to the behaviors expected by or associated with the levels of performance defined in the three-point rating scale — an issue discussed in greater detail in a later section. Last, the current guidelines do not provide teachers with specific examples of evidence that can be used to document their practices and performance related to each particular indicator. This could include, for example, course plans and syllabi, handouts, and additional materials provided to students, examples of student assignments or group projects, examples of feedback provided to the students on written work, and so forth.

One way to tie the teacher evaluation to specific standards of practice is to create a portfolio of the teachers' work in the classroom. Teacher evaluation programs in the United States and elsewhere have used portfolio-based assessments for peer review, as they can provide rich and valid information about teacher practice for longer periods of time than traditional assessments allow.[40]

Criteria for Incorporating Additional Indicators

CM recognizes that even the most extensive list of indicators will not identify every aspect of quality teaching under every circumstance and context. The practice of allowing individual schools to incorporate additional indicators of professional performance based on the local context is useful. However, the current guidelines for incorporating additional indicators may need to be reviewed. The guidelines require that the new indicators address important educational issues and permit observation or documentation. While these two requirements are eminently reasonable, whenever there are disagreements about whether to include a certain indicator, the guidelines require that school personnel "promote a solution that involves all members of the school participating in the program." The precise meaning of this final requirement needs clarification so that it does not prevent the addition of potentially useful indicators that may not be equally relevant to all teachers participating in CM.

Rating Scale

Bias in the distribution of ratings is a common problem in subjective evaluations. An empirical analysis of subjective evaluation in the private sector showed that ratings tend to be compressed and that compression becomes more severe as the ratings become more important for setting pay (Prendergast, 1999, cited in Asch, 2005). Research in education has also shown that principal ratings may also exhibit little variance; that is, most teachers are rated as adequate (Porter, Youngs, and Odden, 2001). One reason could be that

[40] The cases of Chile (*Ministerio de Educación*, 2005) and Douglas County, Colorado (Hall, Caffarella, and Bartlett, 1997) can be cited as examples. See also the Stanford Teacher Assessment Project (Darling-Hammond, 2000).

in systems where supervisors do not derive any personal benefits from a subjective evaluation of their peers or subordinates, they might have an incentive to be lenient so as to minimize conflict and maintain morale (Milgrom and Roberts, 1988, cited in Asch, 2005).

For the past five years (*Etapas* 8 to 12 of CM), the median rating for professional performance has been 9.1. The skewed distribution of the professional performance ratings in CM raises concerns about the usefulness of professional performance in evaluating teacher quality.

The three-point scale currently used to rate professional performance, along with the lack of a clear definition of the desired performance levels, could account in part for the limited variability observed in the ratings. The three points in the scale are only vaguely defined as the "extent and the quality with which the teacher carried out each activity." The instrument lacks an explicit definition of the meaning of each performance level in the rating scale. Without an explicit definition of the performance levels, these are certain to be interpreted differently by members of the committee within a school and by the committees at different schools, which in turn can impact the reliability of the resulting quantitative indicators.

In addition, it should be noted that in the Mexican context the middle of the scale used to grade students (5 out of 10 possible points) has historically constituted a failing grade. Without explicitly labeling each rating level, the committee could well perceive the middle of the 1 to 5 scale (i.e., a rating of 3) as indicating failure. This in turn could result in the committee choosing 5 as the only acceptable alternative – especially since ratings are assigned after deliberation among committee members who may exert peer pressure.

Reliability and Validity

No evidence on the psychometric properties of the peer review instrument was available. In general, surface features of the instrument as currently designed may cause confusion about the appropriate ratings for specific indicators and result in decreased reliability of the measures of teaching practice. An evaluation of the internal consistency and dimensionality (i.e., whether the instrument tests one or various dimensions of knowledge and skills) of the professional performance instrument is needed to determine the extent to which different indicators provide useful information and gauge similar or distinct aspects of teacher practice. However, as currently implemented, the rating and data collection procedures prevent empirical examination of reliability because of the lack of item-level data (only overall points are reported for each teacher at the end of the year) and pilot studies.

In terms of dimensionality, only 6 of 15 indicators in the instrument are directly related to instructional practices or, more generally, the teaching and learning process in the classroom (items 1, 2, 3, 5, 6, and 7). Furthermore, in some cases the distinction between indicators is unclear, while in others, multiple indicators are contained within a single item. As a result, the number of direct indicators of instructional practice involved in evaluating professional performance is limited and their overall quality questionable.

For example, indicator 6 inquires about the extent to which the teacher "adapts instructional strategies and materials and evaluation techniques to the achievement of individual students and the group as a whole." Indicator 7 addresses the extent to which the teacher "implements instructional strategies according to the characteristics of the students to stimulate learning in the group and address individual student needs." As another example, indicator 2 (diagnosis of student knowledge prior to the school year) and the first part of indicator 3 (instructional planning resulting from this diagnosis) are closely related and tap into the same construct. However, the second part of indicator 3 incorporates an entirely different dimension of teacher performance (efforts to involve parents in the education of their children), which is then also the subject of indicator 13.

Peer-Review Process

While the idea of rating by consensus may be appealing in principle, anecdotal evidence and interviews with CM staff indicate that problems of validity have been observed arising from peer pressure by influential members of the evaluation committee. The evidence from a limited number of interviews conducted for this study suggests that for a variety of reasons, the guidelines calling for committee deliberations are often not followed, and principal judgment may be the main and sometimes only factor involved in evaluating the professional performance of each teacher.

Furthermore, it is important to note that, in general, peer-review systems rarely involve teachers evaluating other teachers in the same school building, as is the case in CM. Only one of the peer-review programs we reviewed (in Minneapolis) used this system, and its purposes were essentially formative and geared toward providing assistance for teacher improvement. Results were not tied to incentives or other high stakes decisions. In most other cases, peer reviewers come from other schools in the district or state (in some cases, policies specifically prevent teachers from evaluating their peers in the same school).[41]

One potential issue with CM's peer-review process is that it is summative; that is, it is used only to measure teacher performance for incorporation or promotion into the program. Most peer-review programs in the United States and elsewhere are designed to serve a formative function as well; e.g., to provide assistance for new teachers or teachers performing poorly.[42]

[41] There are studies that suggest that there is a conflict of interest when school principals evaluate teachers because they are playing both an instructional leader and an evaluator role. These authors argue that often principals will choose to avoid conflict by not formally criticizing teachers, resulting in compressed ratings that are often very positive (Wise, Darling-Hammond, Bernstein, and McLaughlin, 1984). These authors, however, do not discuss why this would be the case in education, but not in the private sector; for example, where bosses serve both a leadership (and sometimes a mentoring) and an evaluating function.

[42] The pioneering peer review system in the United States is the Toledo Peer Assistance Plan (PAP), which was the basis for several others, including recent efforts in Columbus, San Francisco, Seattle, and Chicago. The program provides mandatory assistance to new teachers and experienced teachers referred for inadequate performance by their principal. Distinguished teacher consultants are selected by a governing panel and receive professional development to serve as evaluators and mentors, providing information and guidance to help the teacher meet standards of practice defined by the district; consultants also provide a final recommendation that can lead to termination in cases where teachers exhibit serious continued shortcomings. Peer review programs with formative purposes have also been implemented in other countries. In Chile, for example, the Teacher Professional Performance Evaluation System (*Sistema de Evaluación del Desempeño Profesional Docente*) uses peer review and assistance primarily for formative purposes (*Ministerio de Educación*, 2005). Other peer review programs of formative nature include the Cincinnati Public Schools Peer Assistance and Appraisal Program (PAAP), the Rochester Career in Teaching (CIT) program, and the Minneapolis Professional Development System (PDS).

Although the peer-review procedures in CM are intended to serve a formative evaluation purpose, no explicit provisions are made regarding the objectives or procedures to be followed for this formative evaluation, the kinds of feedback or assistance to be provided to teachers, the desired outcomes from the process, or the ways in which future reviews will consider assistance or feedback provided in the past. The procedures merely establish that the "second moment" of evaluation (in the middle of the school year) will "identify dimensions of performance that need to be improved or modified." CM program officials might want to consider reforming the peer-review process to serve a more formative function. There is some evidence to suggest that these kinds of formative exercises can be useful for teachers who want to improve their practice and effectiveness (Manning, 2002; Hertling, 1999; for a contrasting view see Kumrow and Dahlen, 2002).

CONCLUSIONS

Our review of the instruments and procedures used for peer review of teacher professional performance in CM revealed several shortcomings. There is no evidence of the instrument's reliability or validity. The evaluating committee is not presented with explicit teaching standards on which to base subjective judgments about teaching practice. Although it is difficult to comprehensively capture the characteristics of a good teacher, it is possible to delineate a profile of desired skills or competencies that could be based, for example, on the profiles currently used by teacher education institutions in Mexico. Moreover, the limited range of the scale and lack of substantive meaning is worth exploring further.

The peer-review literature suggests that the close link between peers might be problematic as it involves teachers in the same school evaluating each other mainly for summative (and salary) purposes. In fact, CM could take advantage of its horizontal hierarchy system to invite teachers in the highest levels to serve as a new class of mentor teachers. Another possibility is to invite teachers who are near retirement and have performed well in the CM evaluations to receive training and become mentors and evaluators to other teachers. This model would, of course, be more legitimate if the teacher tests truly became a measure of skills and teaching competencies beyond what teachers know about the curriculum.

Finally, this process might be better suited for a formative rather than a summative purpose due to the subjective nature of this factor. The literature reviewed here (particularly in the United States and Latin America) suggests that in most cases these kinds of evaluations are used for formative purposes and are conducted by mentor teachers who have taken a leave from teaching or do not work in the teacher’s school.

6. IMPACT OF CARRERA MAGISTERIAL ON EDUCATIONAL QUALITY

As discussed in the introduction to this report, Carrera Magisterial (CM) has three general goals: (1) help improve the quality of education in Mexico through recognizing and supporting teacher professionalism; (2) improve working, living, and social conditions of teachers; and (3) reward teachers who have the highest performance. In this section we review the available data to assess whether the program's incentives work toward meeting the first of these goals. Most of our analyses use student test scores as the outcome of interest. In some instances, however, we use teacher test scores or peer review ratings as the outcome. In all cases, we make explicit mention of the outcome under consideration.

The first section of this chapter presents an exploratory analysis of the relationship between all the factors measured by CM and student achievement. The relationships between the factors and teacher test scores and peer review ratings are also analyzed. Assessing whether CM's incentives are having a positive impact on educational quality is complicated because of the lack of a clear treatment-control-group experimental design that would allow us to more precisely identify program effects. We implement an alternative approach using a regression-discontinuity design. The analysis first explores whether the existence of the salary bonus had a positive effect on student test scores for teachers attempting incorporation to CM; and, second, whether receiving the salary bonuses (after the teacher has gained incorporation into CM, as well as after any subsequent promotions) has any positive effect on student test scores.

RELATIONSHIP BETWEEN THE PROGRAM'S FACTORS AND STUDENT ACHIEVEMENT AND OTHER OUTCOMES

We begin the impact evaluation by exploring the relationships among the teacher factors measured by the program, and whether these are positively related to student test scores and other outcome variables of interest. If none of the teacher characteristics measured by CM are related to any of these other outcomes, it is unlikely that we would find that the program has any impact on student achievement.

Regression analysis is used to estimate the relationship between teacher characteristics (years of service, educational attainment, teacher test scores, gender) and student test scores for *Etapas* 8 to 12. We also test relationships between teacher characteristics and the teacher test score. The longitudinal nature of the data allows estimation of these relationships within teachers over time. This empirical strategy, using a teacher "fixed effect," is superior to cross-sectional models (that utilize only one year of data), because the focus on changes for individual teachers over time eliminates unobserved fixed factors (e.g., teacher ability) that are likely to be correlated with the outcome of interest. The full methodology used in this analysis is described in Appendix C, which also includes detailed results and descriptive statistics, as well as a description of missing data and teacher participation patterns.

Our analysis shows that, within teachers over time, the relationships among years of service, highest degree earned, teacher test scores, professional development, peer-review ratings and student test scores were weak for both primary and secondary teachers. Results for primary teachers suggest that a one standard deviation change in teacher test scores would be associated with a change in average students' test scores of –0.01 of a standard deviation (see Table C.6 in Appendix C). This result implies no relationship. Similarly, a one standard deviation change in peer-review ratings is associated with an increase of 0.05 of a standard deviation on student test scores. When teachers take a national professional development course, this is associated with an increase in average students' test scores of 0.04 and 0.05 of a standard deviation in the average student and teacher test scores respectively. In all models, undergoing professional development (PD) was more highly associated with the outcome of interest than the actual score obtained on the PD test. The results are very similar at the secondary level, with the exception that undergoing PD is slightly more strongly related to the teacher test score (0.07 of a standard deviation).

In most cases, we found no relationship between a teachers' seniority and her student test scores, teacher test scores, or peer-review ratings. Highest degree earned had few significant relationships with the outcomes we considered. At the primary level, teachers with graduate degrees had slightly lower teacher test scores (professional preparation). At the secondary level, teachers with graduate degrees had slightly higher peer-review ratings (professional performance).[43] Coefficient magnitudes were very small, however. This is probably due to the fact that results are averaged for each teacher over time, and changes in the educational levels of individual teachers over the short period under consideration would be unusual.[44]

The extremely low magnitudes of the relationships suggest that once fixed teacher characteristics (e.g., their ability or motivation) as well as other variables are controlled for, variables that do change over time (e.g., seniority) do not appear to be related to results in the teacher or student tests. The only case in which the relationship is slightly higher (although still below the magnitudes of 0.10 to 0.15 of a standard deviation), is that between undergoing professional development and the score on the teacher test at the secondary level.

Recall from Figure 1.1 that the six factors used by CM (as multidimensional measures of teacher performance) are expected to be significantly related to educational quality. However, these factors do not appear to be related to indicators of educational quality such as student and teacher test scores. In general, the results of this section suggest that the relationships among CM factors are all very weak.

ANALYSIS OF INCENTIVE EFFECTS ON THE YEAR OF INCORPORATION[45]

Teacher incentives are often used to encourage teachers to exert greater effort to improve student outcomes.[46] In fact, one of the assumptions

[43] With respect to student achievement at the secondary level, the model using the total score on the teacher test results in a negative coefficient of having a graduate degree. However, in the model that uses the scores by section of the teacher test, shows a slight positive relationship (of lower magnitude). Therefore, we conclude that the evidence of the relationship between graduate degrees and student achievement in secondary is mixed.

[44] For the regression to work, we need variation in the independent variables. There are very few teachers who changed their highest degree earned (e.g., obtaining a *licenciatura*) in the period between *Etapa* 8 and 12. Therefore, we cannot assess the relationship between education and outcomes during this period.

[45] The methods employed in this section are based on McEwan and Santibañez (2004). See Appendix C for full methodology.

[46] We will remain agnostic on the exact mechanisms by which such an improvement might occur. One possibility is that teachers exert greater effort in the classroom over the course of the school year. Another is that cheating occurs (Jacob and Levitt, 2003); that low-achieving students are excluded from high-stakes testing (e.g., Cullen and Reback, 2002; Figlio and Getzler, 2002); that students receive coaching on test-taking (e.g., Glewwe, Ilias, and Kremer, 2003); or even that schools increase students’ caloric intake on the

presented in Figure 1.1 is that recognizing and supporting teachers and their work yields improvements in educational quality. In this section we examine whether the incentives provided by Carrera Magisterial have induced teachers to improve their students' test scores.

Estimating program effects with CM data is challenging because of the lack of a reasonable comparison group. The key is to compare a group of "incentivized" teachers – or rather, their students' outcomes – to a reasonable comparison group. Unfortunately, Mexican data were collected for the administrative purpose of allocating salary bonuses, and thus include only observations on program participants.

To overcome the limitation of not having a natural comparison group, this analysis uses a regression-discontinuity approach that takes advantage of the fact that, within program participants, not all teachers face the same incentives. The key to this strategy is that some teachers face substantially weaker incentives, and that these teachers are a reasonable counterfactual for the group of teachers facing stronger incentives. The strategy hinges on the fact that there is a well-publicized minimum-point cutoff for incorporation (70 points) and that, of the maximum possible global point score of 100, up to 80 points are distributed to teachers before the student tests are administered.[47] We call this the Initial Point (IP) score. These 80 available points partly reflect input criteria (seniority and education), criteria under the control of the teacher (professional development and teacher test scores), and only one product criteria, the teacher test score (Ornelas, 2002). A significant number of teachers obtain so few points from their IP score that they would not reach the 70-point threshold even if all their students received perfect scores. Other teachers score above the cutoff, even without the additional points from student test scores. Both groups of teachers have a

day of the test (Figlio and Winicki, 2005). For a recent overview of this literature, see Hanushek and Raymond (2002). Similar situations appear to have arisen in Mexico. There have been reports of teachers asking below-average students not to show up to school on the day of the test, or of academically better students being "loaned" amongst teachers (Garcia Manzano, 2004; Ornelas, 2002).

[47] As described previously, the minimum 70-point cutoff is a necessary but not sufficient condition for incorporation. We do not believe this affects the assumptions in the regression discontinuity approach because the cutoff is well publicized (and the final cutoff is not known by the teacher). Therefore, teachers will make their decisions (e.g., to exert more or less effort) based on their probability of reaching the minimum 70-point cutoff.

weak incentive to exert costly effort in an attempt to raise their students' test scores.

A third group of teachers, those with IP scores between 50 and 70 points, face stronger incentives to improve student test scores because they are close to — but not assured of — receiving an award.[48] We argue that the nonlinear structure of the awards introduces a discontinuity in the relationship between a teachers' IP score (excluding points from student test scores) and their student test score. The existence of such a discontinuity could plausibly be attributed to a programmatic effect on student test scores (see full methodology in Appendix C).

Our empirical strategy hinges upon teachers wanting to exert more effort to improve their point score in a factor that only counts for 20 out of a possible 100 points. Some argue that the factor's weight is insufficient to motivate teachers to improve student learning (Ornelas, 2002). Moreover, as was previously discussed, even though one of the program's objectives is to reward the better teachers, the design of the program almost guarantees that a teacher with a university-equivalent degree, who remains in service long enough, undergoes professional development, and obtains a mean score (the mean is around 10 points) on the teacher and student tests, will eventually obtain 70 or more points and achieve incorporation or promotion. While it is true that the factor weight might weaken the power of the incentive to improve student learning, most teachers need the student achievement points to achieve incorporation or promotion. This is particularly the case for new teachers with low IP scores that place them at a disadvantage with respect to more senior or more highly educated teachers. Therefore, the empirical strategy employed in this section is a useful tool for determining program effects.

Because the 70-point cutoff is an important condition for our empirical approach, we begin by assessing whether states adhered to this rule. To do this, we graphed the proportion of teachers obtaining incorporation with point scores higher than 70. The higher this proportion, the more states adhered to the minimum-point score rule.

Figure 6.1 suggests that in *Etapa* 8, of all teachers incorporated or promoted, about 65 percent of secondary teachers and only about 20 percent of

[48] A related phenomenon has been noted in other pay-for-performance systems in which awards are a nonlinear function of a performance measure, such as fixed sales quotas (for a review, see Prendergast, 1999).

primary teachers had total scores higher than 70 points. States adhered much better to the cutoff rule in *Etapa* 9 and beyond. By *Etapa* 12, almost 100 percent of both primary and secondary teacher incorporations went to teachers with scores above the minimum-required cutoff. Table 6.1 shows descriptive statistics for the sample used in this analysis.

The descriptive statistics shown in Table 6.1 suggest that only 16 percent of primary teachers and 28 percent of secondary teachers are in the "strong incentive" group (interval between 50 and 70 points). Almost no teachers have IP scores that place them above 70 points. The vast majority of teachers (84 percent in primary and 72 percent in secondary) have IP scores that place them below 50 points, or in the "weak incentive" group.

Figure 6.1
Proportion of Teachers Promoted or Incorporated with Scores Greater Than or Equal to 70 Points (*Etapas* 8-12)

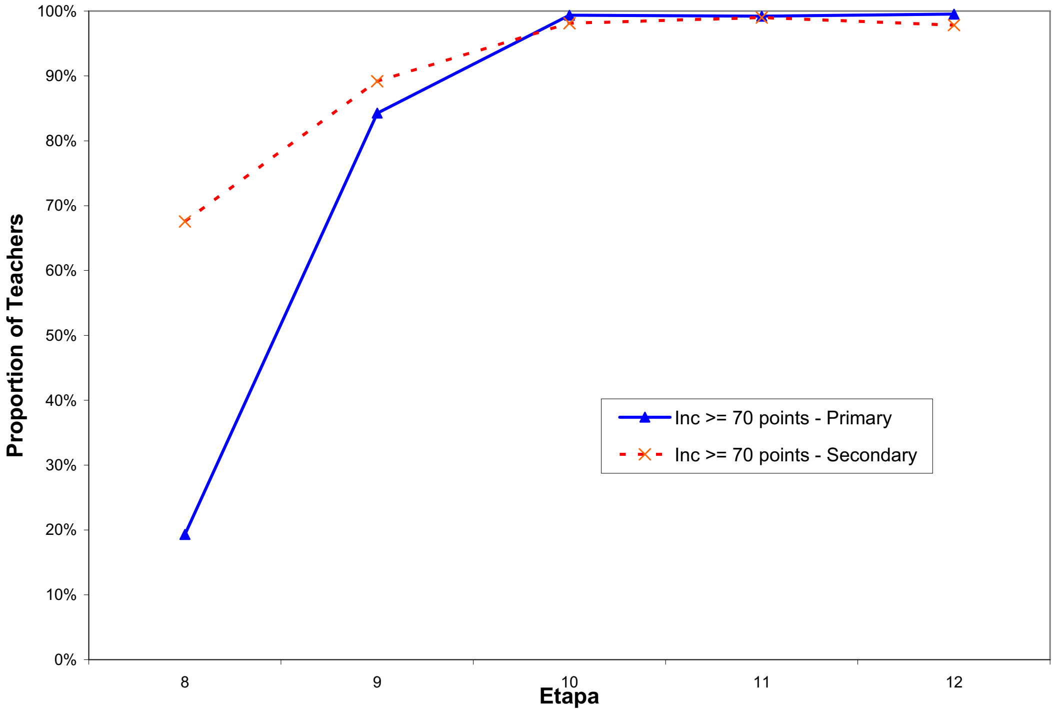


Table 6.1
Descriptive Statistics (Regression-Discontinuity Sample, *Etapas* 9-12)

	Primary School		Secondary School	
Variable	**Mean**	**SD**	**Mean**	**SD**
Raw student test score	37.8	9.0	35.2	9.5
Total point score	53.5	8.8	60.5	8.9
IPs	42.4	7.9	45.8	8.1
50 ≤ IPs < 70	16%	37%	28%	45%
50≤ IPs < 55	9%	29%	14%	35%
55 ≤ IPs < 60	5%	22%	8%	28%
60 ≤ IPs < 65	2%	13%	4%	20%
65 ≤ IPs < 70	0%	5%	1%	10%
IPs ≥ 70	0%	1%	0%	3%

Note A: The number of observations is 108,704 in primary school and 50,526 in secondary school. These observations are almost equally distributed across *Etapas* 9, 10, 11, and 12.
Note B: The student test score is the raw test score (not CM point scores for this factor). Raw student test scores are used as the dependent variable in this analysis (see Appendix C).

Table 6.2 shows estimation results for primary and secondary teachers. Results are only reported for models using state fixed effects, as variation in promotion probabilities across states might affect teacher incentives (results without state fixed effects are similar). Results in columns 1 and 3 subdivide the range of IPs between 50 and 70 into 5-point terms beginning at 50 points and ending at 70 points. These results are for all years under consideration (*Etapas* 9 to 12). Columns three and five show results using only *Etapas* 10 to 12. Because after *Etapa* 10 state adherence to the incorporation cutoff was much more rigorous (see Figure 6.1), this analysis was undertaken to determine the robustness of our model.

Table 6.2
Regression-Discontinuity Results: Incentives During the Year of Promotion

	Primary School		**Secondary School**	
	***Etapas* 9-12**	***Etapas* 10-12**	***Etapas* 9-12**	***Etapas* 10-12**
IPs	0.18***	0.19***	0.08***	0.07***
	(0.007)	(0.007)	(0.007)	(0.008)
50<= IPs<55	-0.08	-0.17	0.18	0.27**
	(0.119)	(0.13)	(0.106)	(0.124)
55<= IPs< 60	-0.16	-0.17	0.14	0.32
	(0.167)	(0.181)	(0.145)	(0.17)
60<= IPs< 65	-0.25	-0.35	0.48**	0.63***
	(0.264)	(0.291)	(0.203)	(0.238)
65<= IPs< 70	-0.17	-0.91	0.53	1.42***
	(0.589)	(0.666)	(0.387)	(0.402)
IPs>= 70	1.73	3.029	0.52	0.43
	(2.152)	(1.759)	(1.022)	(1.023)
Observations	103202	78189	49347	34373
R2	0.15	0.15	0.61	0.62
State fixed effects?	Yes	Yes	Yes	Yes

Note: Robust standard errors, adjusted for school-level clustering, are in parentheses. Other controls include dummies for *Etapas*, states, grades, school shift, municipal economic index, and teacher gender.
***(**) indicates statistical significance at 1% (5%).

All the intervals in Table 6.2 represent teachers in the strong incentive group. To ensure that linearity assumptions of the regression discontinuity design are well supported, the sample used in this analysis includes only teachers with IP scores greater than 30 points.

The coefficients for the IP subdivisions in Table 6.2 suggest no statistically discernible effect of primary teachers being in the strong incentive group on student test scores. At the secondary level, the effect is only statistically different from zero for teachers with IP scores in the interval between 60 and 65 points. Only 4 percent of secondary teachers, however, are in this group (see Table 6.1).

Two kinds of sensitivity analyses were performed to test the robustness of our results. First, we reestimated this analysis only for *Etapa* 10 and beyond, to ensure maximum adherence to the minimum 70-point cutoff rule (see Figure 6.1). The results confirm the nonstatistically significant effect of being in the strong incentive group for primary teachers. If we look only at *Etapas* 10 to 12, we see a positive effect of being in the strong incentive group for secondary teachers with IP scores between 50 and 55 points (14 percent of secondary teachers in the sample), and those between 60 and 65 (4 percent of secondary teachers in the sample), and 65 and 70 points (1 percent of secondary teachers in the sample). The magnitude of these effects is not very large, around 3 to 15 percent of a standard deviation in student test scores.

As a second sensitivity analysis, we excluded teachers with IP scores above 70 points. The results (not shown) remain unchanged, primarily because there are so few teachers with IP scores above 70.

Several reasons could be behind the lack of strong positive effects. First, teachers might not have enough time to affect student test scores. Teachers, particularly those who are new to the program, do not know their total IP score until after they take the teacher test in March. Because the student test is administered in June, even teachers who face strong incentives due to their April IP score might not have enough time during these three months to effect considerable improvements in students' preparedness.

Second, it is possible that even those teachers with strong incentives and reasonable expectations about their IP score do not have the means to effect considerable change on their group of students. Recall from Chapter 2 that one of the assumptions of incentive programs is that teachers know "the technology" to improve student achievement. This is not necessarily the case. In Mexico, educational decision-making in areas such as curriculum, textbooks, and preservice education is highly centralized. And, as discussed in chapter 2, research on accountability in education suggests that teachers might need assistance determining how to improve. Moreover, technical assistance and professional development might not be enough to compensate for insufficient material resources or for the broader context in which some schools operate (Hamilton, 2005).

ANALYSIS OF SALARY BONUS IMPACT AFTER AN INCORPORATION OR PROMOTION

The purpose of the previous analysis was to see what impact, if any, CM incentives had on student test scores as a result of teachers vying for incorporation or promotion. We now take this analysis one step further and ask whether the salary incentives offered by CM had subsequent positive effects after a successful incorporation or promotion into one of CM's higher levels. That is, once a teacher receives the salary bonus associated with an incorporation or promotion, what are the effects on student achievement?[49]

To do this analysis, a first set of regressions was estimated taking advantage of the longitudinal nature of the data and controlling for individual teacher effects. This analysis (results not shown here) resulted in a surprising pattern. Once teachers receive the salary incentive, the probability that their student achievement scores drop in subsequent years is statistically significant, although very modest in magnitude.

These results can be explained in a variety of ways, including the possibility of mean reversion. Many countries report the mean achievement of students in a particular grade in a given school, but these means fluctuate from year to year. Some of these changes are due to real improvements or declines in teacher quality or effort. However, other changes may be due to random shocks or statistical noise (for a discussion, see Kane and Staiger, 2002).[50]

[49] Some of the analyses here use teacher test score and peer review as the outcomes (see Appendix C).

[50] For example, achievement could vary because of one-time shocks such as a schoolwide illness or a poor testing environment. Alternatively, each entering cohort of students is analogous to a random draw from the population of students in a neighborhood. Thus, sampling variation will ensure that mean achievement fluctuates because, in some years, the draw yields a group of better (or worse) students. Sampling variation is likely to be especially pronounced in smaller schools or classrooms (Kane and Staiger, 2002; Chay et al., 2005). Thus, in any given year, mean achievement may be higher (or lower) because of good (or bad) luck. An important point is that such luck — whether good or bad — is unlikely to be repeated in the next year. Thus, a classroom or school that experiences especially high achievement in one year is likely

How might this phenomenon relate to the analysis of CM’s student test scores? Imagine that a group of teachers are seeking incorporation to Level A in a given year. In this group, some teachers will experience particularly good luck, such as a motivated student cohort. Perhaps their students’ performance, in concert with other assessments, is sufficiently high to warrant incorporation. Even if such teachers continue to exert the same level of effort in the subsequent year, their students’ performance — now assessed in a different cohort of students — is unlikely to experience the same positive shock. Thus, it could appear that promotion “causes” achievement to decline, when in fact the decline is merely due to mean reversion.[51]

To assess whether mean reversion affected our results, we reestimated the fixed-effects regressions while excluding the observations of teachers in the year in which they were promoted. For example, consider a teacher with observations in three years: *Etapas* 8, 9, and 10. The teacher was incorporated at the end of year 9; thus, the teacher’s levels are unincorporated (*Etapa* 8), unincorporated (*Etapa* 9), and Level A (*Etapa* 10). The analysis will exclude the middle observation because the teacher may have experienced an especially large positive shock in that year that led to incorporation. The analysis will compare teacher performance in *Etapas* 8 and 10. These results are reported in Appendix C. They suggest that the negative coefficients evidenced were, at least in part, the result of mean reversion.

Because mean reversion could mask the effect of the salary bonus on the various outcomes, we conducted a second empirical strategy that disallows mean reversion. This empirical approach takes advantage of the fact that each of Mexico’s 32 states used an implicit cutoff for incorporation and promotion and uses this fact to implement a regression discontinuity design.[52] As previously discussed, states rank teachers each year according to their total CM point

to experience lower achievement in subsequent years. This phenomenon is a statistical artifact known as mean reversion.

[51] For a similar analysis of a Chilean program, see Chay et al. (2005).

[52] Recall that CM does not have a minimum-point score required for promotion. It does require a minimum score of 70 points for incorporation. This means that teachers who want to become incorporated know about this requirement and (it is assumed) are working toward obtaining this minimum-point score. Teachers who are up for promotion, however, do not know what minimum they should be striving for. When states rank teachers and allocate the budget, they use a cutoff that is determined by their CM budgets that year and the total point scores obtained by teachers. Because the budgets and scores change every year, so do the cutoffs.

score and incorporate or promote teachers until their budget for incorporations and promotions is exhausted. We assume that teachers arbitrarily close to this cutoff are similar in observed and unobserved ways, the only difference being that those who fell just above the cutoff received the treatment (i.e., the promotion and its salary bonus) and those below did not. In regression-discontinuity designs, the program effect is identified not by comparing the outcomes of the treatment and comparison group, but by estimating the change in the prepost relationship between test scores at the cutoff point, thus avoiding mean reversion.[53] In our case, this is done by estimating the change in the relationship between assignment variable (total point score) and student test scores at each of the state cutoff points.

This strategy avoids mean reversion, but it still requires adequate sample retention to maintain internal validity. For example, if teachers who were promoted and return the next year are very different from teachers who do not return, any detected effect might be biased by sample composition. Similarly, if the sample of nonpromoted teachers who return the next year is significantly different than those who were present in the year of assignment into the program, the results might be biased as well.

To assess the direction and magnitude of this bias with the available data, we conducted a descriptive exercise that identifies teachers promoted or incorporated in each *Etapa* (the treatment group) with scores 20 points in either direction of the cutoff and compares their observable characteristics with those who did not return the next year.[54] This analysis is also done on teachers who were not promoted or incorporated in each *Etapa* (i.e., the comparison group). Full results from this analysis can be found in Appendix C. These results suggest that, in most cases, primary teachers who returned were

[53] In regression discontinuity designs we would expect both groups (treatment and comparison) to regress to the mean. In general, we would expect high-scoring teachers to perform worse the year after assignment (when their scores could have experienced a positive shock) and vice versa. This does not pose an internal validity threat because we expect this regression to the mean to be continuous across the range of the assignment variable (or in our case across the range of student test scores and other outcomes). Mean reversion poses a problem for RD only when it results in a discontinuity in the bivariate relationship coincidental with the cutoff point. See Trochim (2006) for a good introduction to RD design.

[54] We also performed this analysis for teachers who were 30 to 100 points around the cutoff. The mean differences in scores are very similar to those we discuss here, although slightly larger.

more experienced, had less education, and higher student test scores than teachers who did not return. This was true regardless of whether teachers returned after a successful or unsuccessful incorporation or promotion attempt. In secondary school, returning and nonreturning teachers were more alike than at the primary level (i.e., the difference between the various scores and teacher characteristics was not statistically different from zero in most instances). However, in the cases where the differences among the two groups were statistically significant, those who returned were slightly more experienced, had slightly less education, and had higher peer review ratings.

These results suggest that sample attrition might be less of a problem at the secondary than at the primary level. Moreover, in most cases, for both primary and secondary teachers, the magnitude of the difference in student and teacher test scores between returning and nonreturning teachers is greater for the nonpromoted group. This means that attrition could affect the estimates through the nonpromoted (comparison) group, more than through the promoted (treatment) group. To the extent that returning teachers in the comparison group have higher test scores this could introduce a downward bias and underestimate any positive effects we find (or overstate any negative effects). Even though most of these differences are not as large as to seriously compromise our estimates, the results from this simple descriptive analysis suggest that sample composition might pose more of a problem for primary teachers. These results, therefore, should be interpreted with caution.

The regression-discontinuity design is implemented by estimating a regression where the outcome in a given *Etapa* (e.g., student test scores) is regressed on the previous *Etapa*'s total point score (i.e., the assignment variable) and a variable for the treatment (whether the teacher was or not promoted or incorporated). If a break is evident near the cutoff each state uses for incorporation and promotion, this should be reflected in the coefficient of the treatment variable. We do this estimation separately for the 32 states and then produce weighted averages of the treatment coefficients. States with very small sample sizes or extreme data distribution are eliminated from the analysis.[55] Results are reported in Tables 6.3 and

[55] This eliminated 29 cases, or 15 percent of the data in the case of primary teachers, and 38 cases, or 20 percent of the data in the case of secondary teachers. However, because most eliminated cases included the

6.4. The coefficients in these tables can be interpreted as treatment effects. A coefficient of −0.76, for example, can be interpreted as the difference in student test scores (attributed to the treatment) between teachers who were incorporated and teachers who were not incorporated into the program. This is equal to an effect size of approximately 8 percent of a standard deviation in student test score (the standard deviation on student test scores is roughly 9 points; see Table 6.1).

Table 6.3
Effects of Incorporation or Promotion on Student Test Scores
(Primary School Teachers)

	Weighted Treatment Coef.[a] (Dep. Var.: Student Test Score)	**Std. Error**	**N**
Teachers who were incorporated	-0.76**	0.303	70
Teachers who were promoted	-3.12**	0.151	93
Teachers who were incorporated or promoted	-2.71**	0.149	163
Teachers who were incorporated in:			
Etapa 9	-0.73	0.442	21
Etapa 10	-0.97**	0.411	26
Etapa 11	-0.57	0.670	23
Teachers who were promoted in:			
Etapa 9	-3.59**	0.276	29
Etapa 10	-3.17**	0.260	32
Etapa 11	-2.50**	0.211	32

Note: The maximum number = 163 cases (representing an average of 810 teaching positions).

[a] The treatment effects we report here are weighted averages over 163 sample cases. Each case represents the effect size (coefficient) from an RD regression in one state and one *Etapa*. On average, each case includes 810 teaching positions (the minimum is 30, the median is 474, and the maximum is 7,703). Original sample included 192 cases, but 29 were deleted because of extremely small sample sizes or extreme data distributions.

** significant at 5%.

smallest states, this elimination brought up the mean number of teachers represented in the sample from 708 to 810 in primary, and 180 to 209 in secondary.

These results suggest that, for primary teachers, there is a negative effect on student test scores after receiving the CM salary bonuses.[56] This effect is small after incorporation (less than ten percent of a standard deviation in student test scores), but much larger after promotion into Level B and beyond (up to 35 percent of a standard deviation on student test scores). Note that biases arising from attrition would suggest that these negative results are somewhat overstated; i.e., they could present an overly pessimistic picture of the effects of incorporation and promotion into higher CM levels (although the available information does not allow us to estimate the magnitude of the bias precisely). Thus, they could be seen as an upper bound of the negative effect.

Table 6.4 shows results for secondary teachers. These results are more robust than those for primary teachers because of better sample sizes by state, and the fact that the normality assumptions hold better (there seems to be less noise in the regression discontinuity plots [not shown here] than in those for primary teachers). The coefficients in Table 6.4 demonstrate, for example, that there is a difference of -0.94 points in student test scores between teachers who were incorporated and teachers who were not incorporated into the program.

The secondary level results also show a pattern of negative effects on student test scores after teachers are incorporated or promoted. The magnitude of the promotion effects, however, is much smaller than was the case for primary teachers (around 10 to 15 percent of a standard deviation in student test scores). This supports our earlier belief that primary level results might be overly pessimistic due to sample composition.

[56] A related phenomenon has been observed in the behavior of salespeople, who may “pull in” sales into one fiscal year, to make a sales quota and earn a bonus, leading to a reduction in sales in the following fiscal year (Oyer, 1998). It should be noted, however, that the product in this kind of activity is much easier to define (and produce) than the “products” of educational activities.

Table 6.4
Effects of Incorporation or Promotion on Student Test Scores (Secondary Teachers)

	Weighted Treatment Coef.[a] (Dep. Var.: Student Test Score)	Std. Error	N
Teachers who were incorporated	-0.94**	0.375	73
Teachers who were promoted	-0.78**	0.257	81
Teachers who were incorporated or promoted	-0.83**	0.212	154
Teachers who were incorporated in:			
Etapa 9	-0.19	0.905	21
Etapa 10	-1.40**	0.559	25
Etapa 11	-1.19**	0.479	27
Teachers who were promoted in:			
Etapa 9	-1.60**	0.430	25
Etapa 10	-0.88	0.470	29
Etapa 11	0.34	0.350	27

Note: The maximum number = 154 cases (representing an average of 209 teaching positions).
[a] The treatment effects we report here are weighted averages over 154 sample cases. Each case represents the effect size (coefficient) from an RD regression in one state and one *Etapa*. On average, each case includes 209 teaching positions (minimum is 21, median is 142, maximum is 1,335). Original sample included 192 cases, but 38 were deleted due to extremely small sample sizes or extreme data distributions.
** significant at 5%.

CONCLUSIONS

Overall, the results from this section suggest that CM's incentives do not have any discernible effects on student test scores for primary teachers who are vying for incorporation. The program's incentives appear to have modest positive effects (on the order of 3 to 15 percent of a standard deviation) on student test scores for secondary teachers who are vying for incorporation and are in the strongest incentive groups. These effects are most apparent after *Etapa* 10 (when states adhered more closely to the minimum cutoff required for incorporation) and are only evident for about 20 percent of the secondary teachers in the sample. The largest effects (those of around 15 percent of a standard deviation on student test scores) are only apparent for 1 percent of the secondary teachers in the sample.

The incentives to improve student test scores decrease significantly once teachers achieve incorporation or promotion. The negative effects of incorporation and promotion into CM on student test scores appear to be larger after promotions into Levels B and beyond. These results can be explained in part by sample attrition, but they could also be the result of CM’s teacher incentive structure. Because teacher salary bonuses are guaranteed for one’s entire career, teachers might have less of an incentive to exert additional effort after receiving a promotion. Once promoted, teachers may return to their normal effort level in the following year, or may perhaps even reduce effort given that their promotion is binding. The economic literature on incentives (see chapter 2) provides some evidence to suggest that the use of repeated incentives at various points in time (as opposed to incentives based primarily on one-shot gains) might improve efficiency by reducing risk to the worker of having a bad year as well as the possibility of promoting teachers that had an unusually good year but whose normal performance levels are actually average or below average.

Overall, the findings from this section point to a consistent underlying explanation for the impact of CM: the program has little discernible effect on student test scores for most teachers (assuming they have reasonable information about their own abilities); those few teachers who react favorably to CM’s incentives appear to work harder to improve their students’ test scores only when it matters.

Last, it is worth noting that the analyses in this chapter use CM’s own measures of educational quality (student and teacher test scores, peer-review ratings) as variables of interest. To the extent that these variables do not accurately reflect educational quality, the results shown here will suffer from the same limitation.

The findings in Chapter 4 relating to the technical properties of the teacher and student tests suggest that in some cases neither the tests themselves nor the assessment procedures are adequate and that the tests mainly assess low-level skills. Inadequate test validity and reliability do not directly affect the empirical strategy used in this paper, however. Even if using flawed measures of student achievement or teacher knowledge, we should observe that the teachers in the strong incentive group differ on average from teachers in the weak incentive group.

Perhaps one lesson to be drawn from the findings regarding the low levels of cognitive skills demanded by the teacher tests coupled with the low relationship between teacher test results and student achievement is that these tests could be improved in ways that come closer to measuring teaching competencies. Such assessments could include not just knowledge of the curricula, but an assessment of teachers' basic literacy and analytic and critical skills, as well as some provision for a more in-depth evaluation of their teaching practice through interviews or observations. Those assessments could at the very least focus on a much broader definition of what constitutes teacher ability or knowledge than what is currently measured with the teacher test.

In our analyses, the results concerning the effect of the program's incentives on student achievement should be interpreted cautiously as effects on student test scores in the CM tests and not as effects on student learning in general.

7. POLICY RECOMMENDATIONS AND FINAL CONSIDERATIONS

Carrera Magisterial (CM) is one of the pioneer teacher incentive programs in the world. Mexico was one of the first countries to explicitly tie the salary of all public school teachers to teacher performance. However, the political nature of the program, run centrally by joint commission of Ministry of Education and teacher union officials (*Comisión Nacional* SEP-SNTE), in which the main actors do not always agree on a common vision, has produced ambiguities with respect to program objectives. CM could be described as a "middle of the road" program — falling somewhere between merit pay (SEP's vision) and a new salary schedule (SNTE's vision).

Perhaps owing to this tension, while other teacher incentive programs are often criticized for their overreliance on student test scores as the sole measure of teacher performance, CM adopts a broader definition of performance. This definition includes not only student achievement, but also the knowledge and skills acquired through professional development and educational attainment, seniority, and peer and supervisor evaluations. CM gives teachers the opportunity to demonstrate their knowledge and teaching competencies through both objective and subjective measures. Its assessment process is largely of no cost to teachers, and the salary bonuses should be, in theory, large enough to provide strong incentives to perform well.

The program, however, hasn't always been implemented as designed. Owing perhaps to competing agendas, more than half the teachers currently a part of CM were automatically enrolled in the program during its first two years. These teachers obtained salary bonuses without having to undergo a formal assessment. Furthermore, the results from this analysis suggest that both the objective (teacher and student test) and the subjective (peer ratings) measures used to gauge teacher performance might contain significant flaws. Moreover, the incentives offered by the program (salary bonuses associated to incorporation or promotion) show, from a statistical point of view, zero or very modest positive effects on student achievement.

The purpose of this study was to evaluate the adequateness of CM's instruments for teacher assessment as well as the impact of the salary incentives on some indicators of educational quality. This section describes some general policy implications (as well as some reforms of a more structural nature) that might help CM improve its assessment system.[57] These policy implications highlight the challenges associated with implementing and administering such large scale teacher incentive programs, and should serve as a guide to policymakers in Mexico and other countries.

Before discussing this report's recommendations, some limitations to the analysis should be acknowledged. First, this research uses CM's own measures of quality as outcomes of interest (student test scores, teacher test scores, and peer-review ratings). To the extent that these measures do not accurately reflect educational quality, our results will suffer from the same shortcoming. Second, findings regarding the sometimes inadequate technical properties of the tests, the low levels of cognitive demand some of the tests elicit, and other considerations having to do with the assessment process, force us to adopt a restricted interpretation of educational quality. The results of the impact of CM's incentives on student achievement should be interpreted not as the impact of the program on student learning, but as the impact on student performance on the CM tests.

This restricted interpretation does not undervalue the analysis because the program itself uses these measures to reward teachers. In addition, these analyses represent what is perhaps the best possible empirical estimation that can be performed with the available data. The nature of the CM program and the information it collects restrict the methodological choices. Therefore, some methodologically superior strategies such as natural experiments or analysis using a baseline from which to measure improvements in learning were not possible. Some of the most important data limitations have to do with voluntary teacher participation, data collected mainly for administrative purposes with no information on nonparticipating teachers, teachers coming in and out of the dataset at various points, incomplete information in some cases, and lack of information about students and schools, among others.

RECOMMENDATIONS

SEP expends a considerable amount of effort testing millions of students every year. These testing efforts currently absorb more than 90 percent of the budget allocated to the Ministry of Education's Directorate of Evaluation

[57] A list of more detailed policy recommendations was prepared for the Mexican Ministry of Education as a project memorandum.

(DGE).[58] The challenge of evaluating millions of students and thousands of teachers using up to 45 different tests, some with multiple forms, is daunting. The fact that the Ministry (through DGE and state authorities) is able to administer these tests every year with a limited budget is remarkable. Perhaps owing to this lack of resources, the results from this report suggest important shortcomings in the quality assurance in CM's testing program, which result in significant technical limitations (e.g., poor psychometric properties and low reliability in some cases). The main recommendations in this report relate to the shortcomings observed in the assessments and assessment procedures used by CM to determine incorporation and promotion.

Steps can be taken to ensure that the instruments used by test-based accountability systems, such as CM, meet internationally accepted standards of quality. Technical manuals for development, administration, and reporting of teacher and student tests should be created and regularly updated. Manuals could incorporate quality assurance provisions for tests during the test development phase, prior to administration, and after administration. These manuals might also include provisions to review items for different kinds of biases (e.g., gender, cultural, regional). They should also include detailed guidelines for confidentiality and communication of test results and other information to teachers, schools, state authorities, researchers, and the public. New tests should be piloted. If time and budgetary constraints prevent item piloting, quality control in earlier stages of test development (including test specifications, item writing, revision, and analysis) acquires particular importance.

Another issue to consider is that, contrary to what happens in other countries, the teacher test measures only the knowledge teachers have over the curriculum they teach (as well as teaching methodologies and legal aspects of the Mexican education system). In addition, the CM tests do not become more difficult as teachers seek promotion into higher levels even though these levels are associated with higher salary bonuses. Policymakers might consider

[58] For example, DGE's operative budget in 2004 was between eight and nine million dollars. This is not very high compared with INEE's budget, which was around 12 million dollars, and to other school improvement programs such as *Enciclomedia* (around 200 million). For more budget information, see http://www.shcp.sse.gob.mx/contenidos/presupuesto_egresos/temas/pef/2005/temas/tomos/11/r11_afpe.pdf.

developing a test that would measure both subject-matter knowledge as well as individual and teaching competencies or skills. This test would seek to identify teachers with better abilities both to receive the higher incentives and to serve as potential mentors or evaluators of other teachers. To undertake this more in-depth evaluation, the teacher tests would not be multiple choice only but could include open questions, class or instruction profile developments, and interviews.[59]

Other recommended reforms are more structural in nature. The results from this study regarding the impact of the incentives offered by the program on student achievement suggest that such impact is either zero or very small, and that it can become negative once the incentive has been received (although the magnitude of this latter effect is not very large). Therefore, policymakers might consider reviewing the main features of CM's assessment system, the factors it includes, and how these are assessed. For example, incentive programs should avoid "double-counting." If seniority and education are already determining the base salary, they should not also determine the size of the bonus. This is particularly so when these have not been found to be strongly linked to the outcomes of interest (as is the case in CM).

Policymakers may also want to consider assessment measures that cover more than one year's worth of data, with the objective of providing continuous incentives to improve. This could have a two-fold effect: (1) it may provide incentives for teachers to exert high levels of effort each year; and (2) the averaging of performance across years would reduce the amount of random noise in test scores, thus providing a more statistically precise measurement of teacher performance. Policymakers can also consider the use of penalties if performance drops below an acceptable level.

Incentive programs should also discourage gaming behaviors and reduce the influence of noise or other factors not related to the outcomes of interest.

[59] One possibility is to substitute the legal aspects section of the test (which has very low reliability and little empirical support) with a section on basic skills or competencies. For tests used to determine incorporation, this might be done using as models initial certification tests that assess basic math and reading and writing skills (such as the Praxis I tests discussed in Chapter 4). As teachers progress in CM, this section could be modified to be more difficult. Another option is to substitute the teaching methodologies section of the test (also with low reliability) with a basic skills one if the peer-review and professional development factors can be modified to more accurately capture information about teaching practice.

For example, if one of the outcomes of interest is teacher-induced improved student achievement, the program should ensure that student achievement can be attributed to the teacher. This can be done using value-added methods. Such methods purport to provide precise and valid estimates of the contribution of teachers to the gains in the achievement of their students, which take into account the socioeconomic composition of the classroom (McCaffrey, Koretz, Lockwood, and Hamilton, 2004).[60] A second less exhaustive and data-intensive option (although superior to what is currently done) is to collect student background measures such as socioeconomic status and then compare teachers to others with similar students. Even though CM adjusts its student achievement scores by the socioeconomic status of the region, this might not be sufficient in accounting for differences in student background or other characteristics across schools. Large regions such as Mexico City have geographically proximal schools with very different populations in average socioeconomic status.

Assessment programs such as CM should consider establishing an independent technical and content advisory board to oversee key psychometric and statistical aspects of the tests, their content coverage, as well as compliance with necessary documentation, security, fraud protection, and other key aspects of assessment processes. Such a board would ensure continuous quality assurance of the tests and the testing process.

The peer-review factor could also be adjusted so that teachers do not evaluate others in the same school. This practice might be associated with the low variability in the ratings.

CM has placed a strong emphasis on professional development. But it appears that taking national professional development courses is only weakly

[60] It is important to note, however, that value-added approaches carry with them important requirements in testing and data management. For example, they assume annual testing of students using tests that are vertically equated (to create a scale that allows for measurement of student growth) and the presence of datasets that link teachers to students and track individual student scores across years and grades. Testing students at the beginning and end of the school year would be ideal for estimating teacher value-added effects, as it provides a direct measurement of the initial and final achievement level of each teacher’s classroom. However, testing students twice is usually not feasible, as it effectively doubles the already considerable costs of testing and the demands of time and resources for schools and students. Where annual testing systems are in place, using the results from the previous year as a starting point for estimating teacher effects is often used as an acceptable alternative.

related to improvements on the teacher test, the student test, or peer-review ratings. Even at the secondary level, where the relationship between peer-review ratings and undergoing professional development is higher, the magnitude of the relationship is not very large. The literature suggests that teachers might need assistance in determining optimal strategies for improving student achievement. This is particularly true of teachers working in more disadvantaged settings. Although evaluation of the professional development factor, its instruments, and course content fell outside the scope of this work, our findings suggest that this factor warrants further consideration.

Policymakers might also consider alternate uses of processes already in place.[61]

CM's peer-review component, for example, could serve a formative function instead of being used only to rate teachers' ultimate performance. Peer review as a formative exercise could take place early in the school year to establish clear goals for teaching practice (based on comprehensive teaching standards). At the end of the year, using the available documentation of teacher performance (e.g., student test scores, teacher portfolios, surveys of parents and students, classroom observations), the peer-review committee or the teachers' supervisor could evaluate and deliberate on the teachers' performance with respect to the goals set at the beginning of the year. This kind of performance evaluation is reflective of international practice. Another possibility is for teachers in the highest levels of CM to serve a mentoring and evaluating function. The legitimacy of this option, however, would depend on improvements to CM and its assessment system so that teachers in the highest levels are those that have demonstrated greater teaching competencies.

Teacher incentive programs, as any public policy program, should strive to use resources more efficiently. This is particularly important when, as is the case in Mexico, resources are scarce. For example, the assessment could be targeted to those teachers who actually have strong incentives to perform. In the case of CM, teachers whose initial point scores leave them little or no

[61] This should not be confused with using data results for purposes other than those they were intended to serve, a practice which should be avoided (see chapter 4). If the instruments want to be used for other purposes (e.g., ranking the "best" schools in the country over time or comparing CM results with those of other national tests), they should be reported with all the relevant documentation supporting the appropriateness of such practices.

possibility of being incorporated in a given year (i.e., their final point score would not place them above 70, even with perfect teacher and student test scores) are part of the “weak-incentive” group and should probably not be tested. Testing thousands of these teachers and their students diverts valuable resources from conducting better assessments of teachers facing stronger incentives. Alternatively, given that CM already tests about half its teacher (and consequently student) population every year, an alternative could be to test students every other year, regardless of whether the teacher is enrolled and/or eligible for promotion that year, and use an average of repeated testing as the basis for assigning CM points for student achievement.

Efficiencies can also be gained by coordinating databases across departments. In the case of CM, because failure to undergo professional development is one of the largest causes of teachers obtaining zero total scores, the evaluation authorities should know in advance which teachers have not completed this factor so that they can be excluded from further testing (DGE already does this for teachers who do not take the teacher test).

Last, program impacts are observed using measurements that the CM program has developed to represent educational quality. Even the best-designed incentive system might not show improvements in educational quality if the instruments used to measure it are flawed. If the instruments used by the program do not correspond to what educational authorities, teachers’ unions, researchers, students, and parents understand to be valid measures of student achievement and good teaching practices, the program will not reward what the community in general sees as educational quality.

SCOPE OF THE RESEARCH, DIRECTIONS FOR FUTURE RESEARCH, AND FINAL CONSIDERATIONS

In countries like Mexico, where a national curriculum is in place, educational authorities could focus their resources on producing a single, high-quality national test that could be used by CM as well as by other programs that seek to measure student achievement. The Ministry of Education could then focus its efforts and resources currently spent administering the CM tests and other tests like the *Prueba Nacional Estándares*, on creating a comprehensive assessment system that can serve multiple purposes, including generating value-added measures for teachers and students.

It would be desirable for policymakers to review the objectives, characteristics, and requirements of the assessment systems currently in place to ensure that they complement each other and are aligned with a national assessment policy. Establishing a national assessment policy with concomitant assessment strategies should be based on solid research to inform decision-makers on the advantages and disadvantages of implementing various assessment schemes.[62]

Because they fall outside this study's scope of work, some relevant questions remain unanswered. Our study was limited to primary and secondary teachers who in CM are referred to as *primera vertiente*. The findings of this report apply only to these teachers. A further evaluation of the impact of the program on school administrators' and teacher aides' outcomes is important if CM officials want to make improvements affecting the evaluation for these two groups of education professionals.

Student and teacher test analyses were limited to primary students and teachers and math, Spanish, geography, history, and civics and ethics students and teachers at the secondary level. Other secondary tests, such as chemistry, physics, English, French, and biology were not analyzed. Neither did we evaluate the instruments used to test for knowledge of the various professional development courses. Last, the lack of external measures of learning limited our ability to determine whether CM's student and teacher tests were externally valid. Future research collecting these kinds of data is warranted.

We provide three other suggestions for future research here. We suggest first that future research focus on whether longitudinal measures of the "value-added" of teachers could be incorporated into the CM model. Second, future studies could explore the impact of CM on other dimensions of educational quality that are important to the community, including external measures of learning (e.g., measures not produced by the program itself). A third interesting avenue for future research, also beyond the scope of this report, could be the impact of teacher incentive programs in the context of other policies or interventions to improve education quality, such as the

[62] SEP's current efforts and advances toward a national evaluation system (using national censal exams) are an opportunity in this regard.

interaction between CM and other education reform programs in Mexico like the quality schools program (*Programa Escuelas de Calidad*).

It is clear that the research on large-scale teacher incentive and assessment programs, of which this study is but a small part, is preliminary at best, with many outstanding questions. In general, the results of this research question whether individual teacher incentive programs are the best choice to improve teacher quality and student achievement. It is possible that individual salary incentive programs for teachers, like CM, have less potential than models described as successful in the business and economics literature (often for programs focused on less complex products or activities). Even though the education literature does present cases where teacher incentive programs have resulted in student achievement gains, most of the programs are small scale and their results are small and often short-lived. In some cases, these results were not necessarily the product of increased teacher effort but were due to gaming or other such mechanisms. This suggests the possibility of considering the introduction of other kinds of incentives (group, combined, etc.), although as we discussed in chapter 2, the evidence on those types of incentives is also mixed.

REFERENCES

Achieve, Inc., *Ready or Not: Creating a High School Diploma That Counts,* Washington, DC: American Diploma Project, 2004.

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

Asch, Beth, "The Economic Complexities of Incentive Reforms," in Robert Klitgaard and Paul C. Light, eds., *High-Performance Government: Structure, Leadership, Incentives,* Santa Monica, CA: RAND Corporation, MG-256-PRGS, 2005.

Asch, Beth, and Lynn Karoly, *The Role of the Job Counselor in the Military Enlistment Process*, Santa Monica, CA: RAND Corporation, MR-315-P&R, 1993.

Ávila, Carrillo E., and Brizuela H. Martínez, *Historia del Movimiento Magisterial (1910–1989): Democracia y Salario,* México, D.F.: Ediciones Quinto Sol, 1990.

Bloom, Benjamin, M. Englehart, E. Furst, W. Hill, and David Krathwohl, *Taxonomy of Educational Objectives: The Classification of Educational Goals. Handbook I: Cognitive Domain,* New York, NY: Longmans, Green, 1956.

Brennan, Robert L., "The Conventional Wisdom About Group Mean Scores," *Journal of Educational Measurement*, Vol. 32, No. 4, 1995, pp. 385–396.

Brewer, Dominic J., "Career Paths and Quit Decisions: Evidence from Teaching," *Journal of Labor Economics,* Vol. 14, No. 2, 1996, pp. 313–339.

Chay, Kenneth Y., Patrick J. McEwan, and Miguel Urquiola, "The Central Role of Noise in Evaluating Interventions that Use Test Scores to Rank Schools," *American Economic Review,* Vol. 95, No. 4, 2005, pp. 1237–1258.

Comisión Nacional SEP-SNTE, *Lineamientos Generales de Carrera Magisterial,* Ciudad de México, México: Secretaría de Educación Pública, 1993.

———, *Lineamientos Generales de Carrera Magisterial,* Ciudad de México, México: Secretaría de Educación Pública, 1998.

———, *Normas y Procedimiento para Evaluar el Factor Desempeño Profesional,* Ciudad de México, México: Secretaría de Educación Pública, 2000.

Coordinación Nacional de Carrera Magisterial, *Carrera Magisterial. Antología,* México D.F., México: Secretaría de Educación Pública, 2001.

Cronbach, Lee J., and Richard J. Shavelson, "My Current Thoughts on Coefficient Alpha and Successor Procedures," *Educational and Psychological Measurement*, Vol. 64, No. 3, 2004, pp. 391–418.

Cronbach, Lee J., Robert L. Linn, Robert L. Brennan, and Edward Haertel, Generalizability Analysis for Performance Assessments of Student Achievement or School Effectiveness, *Educational and Psychological Measurement,* Vol. 57, No. 3, June 1997, pp. 373–399.

Cullen, Julie B., and Randall Reback, *Tinkering Toward Accolades: School Gaming Under a Performance Accountability System,* unpublished manuscript, University of Michigan, 2002.

Darling-Hammond, Linda, "Teacher Quality and Student Achievement: A Review of State Policy Evidence, *Education Policy Analysis Archives*, Vol. 8, No. 1, 2000.

———, "Standard Setting in Teaching: Changes in Licensing, Certification, and Assessment," in Virginia Richardson, ed., *Handbook of Research on Teaching,* Washington, DC: American Educational Research Association, 2001, pp. 751–776.

Dee, Thomas S., and Benjamin J. Keys, "Does Merit Pay Reward Good Teachers? Evidence from a Randomized Experiment," *Journal of Policy Analysis and Management,* Vol. 23, No. 3, 2004, pp. 471–488.

Dethlefs, Theresa M., Vickie Trent, Robert M. Boody, Gene M. Lutz, Vicki Robinson, and William Waack, *Impact Study of the National Board Certification Pilot Project in Iowa,* Des Moines, IA: *Iowa* State Department of Education, 2001.

Dirección General de Evaluación (DGE), *Programa de Carrera Magisterial: Factor Aprovechamiento Escolar. Informe de Resultados Nacional. Evaluación 1999-2004,* Ciudad de México, México, 2004.

Dwyer, Carol A., "Psychometrics of *Praxis III*: Classroom Performance Assessments," *Journal of Personnel Evaluation in Education,* Vol. 12, No. 2, 1998, pp. 163–187.

Education Trust, *Ticket to Nowhere: The Gap Between Leaving High School and Entering College and High-Performance Jobs,* Washington, DC: Education Trust, 1999.

Ezpeleta, Justa, and Eduardo Weiss, "Las Escuelas Rurales en Zonas de Pobreza y Sus Maestros: Tramas Preexistentes y Políticas Innovadoras," *Revista Mexicana de Investigación Educativa*, Vol. 1, No. 1, 1996, pp. 53–69.

Figlio, David N., "Can Public Schools Buy Better Qualified Teachers?" *Industrial and Labor Relations Review,* Vol. 55, 2002, pp. 686–699.

Figlio, David N., and Lawrence S. Getzler, *Accountability, Ability, and Disability: Gaming the System,* Working Paper No. 9307, Cambridge, MA: National Bureau of Economic Research, 2002.

Figlio, David N., and Joshua Winicki, "Food for Thought: The Effects of School Accountability Plans on School Nutrition," *Journal of Public Economics,* Vol. 89, 2005, pp. 381–394.

García Manzano, María del Socorro, *Una Mirada al Esquema de Carrera Magisterial en México: Recorrido Histórico y Estudio de Caso en Dos Escuelas Primarias del D.F.,* thesis, México D.F., México: Departamento de Investigaciones Educativas, Cinvestav, 2004.

Glewwe, Paul, Nauman Ilias, and Michael Kremer, *Teacher Incentives,* Working Paper No. 9671, Cambridge, MA: National Bureau of Economic Research, 2003.

Gritz, R. Mark, and Neil Theobald, "The Effects of School District Spending Priorities on Length of Stay in Teaching," *Journal of Human Resources,* Vol. 31, No. 3, 1996, pp. 477–512.

Guarino, Cassandra, Lucrecia Santibañez, Glenn A. Daley, and Dominic J. Brewer, *A Review of the Research Literature on Teacher Recruitment and Retention,* Santa Monica, CA: RAND Corporation, TR-164-EDU, 2004. Online only: http://www.rand.org/publications/TR/TR164.

Hall, Gene, Edward Caffarella, and Ellen Bartlett, "Assessing Implementation of a Performance Pay Plan for Teachers: Strategies Findings, and Implications," Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL, March 1997.

Hamilton, Laura S., "Lessons from Performance Measurement in Education," in Robert Klitgaard, and Paul C. Light, eds., *High-Performance Government: Structure, Leadership, Incentives,* Santa Monica, CA: RAND Corporation, MG-256-PRGS, 2005.

Hamilton, Laura S., Brian M. Stecher, and Stephen P. Klein, eds., *Making Sense of Test-Based Accountability in Education*, Santa Monica, CA: RAND Corporation, MR-1554-EDU, 2002.

Hannaway, Jane, "Higher Order Skills, Job Design, and Incentives: An Analysis and Proposal," *American Educational Research Journal,* Vol. 29, No. 1, 1992, pp. 3–21.

Hanushek, Eric A., "Outcomes, Costs, and Incentives in Schools," in Eric A. Hanushek, and David W. Jorgenson, eds., *Improving America's Schools: The Role of Incentives,* Washington, DC: National Academy Press, 1996.

———, "Assessing the Effects of School Resources on Student Performance: An Update," *Educational Evaluation and Policy Analysis*, Vol. 19, No. 2, 1997, pp. 141–164.

Hanushek, Eric A., and Margaret E. Raymond, "Improving Educational Quality: How Best to Evaluate Our Schools?" in Y. K. Kodrzycki, ed., *Education in the 21st Century: Meeting the Challenges of a Changing World,* Boston, MA: Federal Reserve Bank of Boston, 2002.

Hertling, Elizabeth, "Peer Review of Teachers," *Eric Digest,* Vol. 126, May 1999. Online at http://www.ericdigests.org/1999-4/peer.htm.

Holmstrom, Bengt, and Paul Milgrom, "Multitask Principal-Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design," *Journal of Law, Economics, and Organization*, Vol. 7, special issue, 1991, pp. 24–52.

Instituto Nacional para la Evaluación Educativa (INEE), *La Calidad de la Educación Básica en México: Primer Informe Anual.* México D.F., México: INEE, 2003.

Jacob, Brian A., and Steven D. Levitt, "Rotten Apples: An Investigation of the Prevalence and Predictors of Teacher Cheating," *Quarterly Journal of Economics,* Vol. 118, 2003, pp. 843–878.

Joint Committee on Standards for Educational Evaluation (Joint Committee), James R. Sanders, Chair, *The Program Evaluation Standards. How to Assess Evaluations of Educational Programs, 2nd Edition,* Thousand Oaks, CA: Sage, 1994.

Kane, Thomas J., and Douglas O. Staiger, "The Promise and Pitfalls of Using Imprecise School Accountability Measures," *Journal of Economic Perspectives,* Vol. 16, No. 4, 2002, pp. 91–114.

Kelly, Philip P., "A Comparative Analysis of Teacher Peer Review Programs in Four Urban Districts: Professional Unionism in Action," Paper Presented at the annual meeting of the American Educational Research Association, San Diego, CA, April 13–17, 1998.

Kirby, Sheila, Mark Berends, and Scott Naftel, "Supply and Demand of Minority Teachers in Texas: Problems and Prospects," *Educational Evaluation and Policy Analysis,* Vol. 21, No. 1, 1999, pp. 47–66.

Klerman, Jacob A. "Measuring Performance," in Robert Klitgaard, and Paul C. Light, eds., *High-Performance Government: Structure, Leadership, Incentives,* Santa Monica, CA: RAND Corporation, MG-256-PRGS, 2005.

Koretz, Daniel M., "Limitations in the Use of Achievement Tests as Measures of Educators' Productivity," *Journal of Human Resources*, Vol. 37, No. 4, 2002, pp. 752–777.

Kumrow, David, and Becky Dahlen, "Is Peer Review an Effective Approach for Evaluating Teachers?" *Clearing House,* Vol. 75, No. 5, 2002, pp. 238–241.

Lavy, Victor, "Evaluating the Effect of Teachers' Group Performance Incentives on Pupil Achievement," *Journal of Political Economy,* Vol. 110, 2002, pp. 1286–1317.

Lazear, Edward P., "Why Is There Mandatory Retirement?" *Journal of Political Economy*, Vol. 87, 1979, pp. 1261–1264.

———, "Pensions As Severance Pay," in Zvi Bodie and John Shoven, eds., *Financial Aspects of the United States Pension System*, Chicago, IL: University of Chicago Press, 1983, pp. 57–89.

———, "Performance Pay and Productivity," *American Economic Review*, Vol. 90, No. 5, 2000, pp. 1346–1361.

Loeb, Susanna, and Marianne Page, "Examining the Link Between Teacher Wages and Student Outcomes: The Importance of Alternative Labor Market

Opportunities and Non-Pecuniary Variation, *Review of Economics and Statistics,* Vol. 82, No. 3, 2000, pp. 393–408.

Manning, Renfro, *The Teacher Evaluation Handbook: Step-by-Step Techniques and Forms for Improving Instruction,* San Francisco, CA: Josey-Bass, 2002.

Martinez-Rizo, Felipe, Eduardo Backhoff Escudero, Sandra Castañeda Figueiras, Arturo de la Orden Hoz, Sylvia Schmelkes del Valle, Guillermo Solano Flores, Agustín Tristán López, y Rafael Vidal Uribe, *Estándares de Calidad para Instrumentos de Evaluación Educativa,* Ciudad de México, México: Consejo Asesor Externo, CENEVAL, 2000.

McCaffrey, Daniel F., Daniel Koretz, J. R. Lockwood, and Laura S. Hamilton, *Evaluating Value-Added Models for Teacher Accountability*, Santa Monica, CA: RAND Corporation, MG-158-EDU, 2004.

McEwan, Patrick J., and Lucrecia Santibañez, “Teacher Incentives and Student Achievement: Evidence from a Large-Scale Reform in Mexico,” unpublished manuscript, 2004. Online at http://emlab.berkeley.edu/users/webfac/chay/e251_s05/mcewan.pdf.

Ministerio de Educación, *Instructivo General: Sistema de Evaluación del Desempeño Profesional Docente 2005,* Santiago, Chile, 2005.

Mizala, Alejandra, and Pilar Romaguera, “Regulación, Incentivos y Remuneraciones de los Profesores en Chile,” en C. Cox, ed., *Políticas Educacionales en el Cambio de Siglo: La Reforma del Sistema Escolar de Chile,* Santiago, Chile: Editorial Universitaria, 2003.

Murnane, Richard, and David Cohen, “Merit Pay and the Evaluation Problem: Why Most Merit Pay Plans Fail and a Few Survive,” *Harvard Educational Review,* Vol. 56, February 1986, pp. 1–17.

National Research Council (NRC), *Testing Teacher Candidates: The Role of Licensure Tests in Improving Teacher Quality,* Committee on Assessment and Teacher Quality, K. J. Mitchell, D. Z. Robinson, B. S. Plake, and K. T. Knowles, eds., Board on Teaching and Assessment, Washington, DC: National Academy Press, 2001.

Ornelas, Carlos, “Incentivos a los Maestros: La Paradoja Mexicana,” in Carlos Ornelas, ed., *Valores, Calidad y Educación.* México D.F., México: Santillana/Aula XXI, 2002.

Ortiz Jiménez, Maximino, *Carrera Magisterial: Un Proyecto de Desarrollo Profesional,* Cuadernos de Discusión 12, México D.F., México: Secretaría de Educación Pública, 2003.

Oyer, Paul, “Fiscal Year Ends and Nonlinear Incentive Contracts: The Effect on Business Seasonality,” *Quarterly Journal of Economics,* Vol. 113, No. 1, 1998, pp. 149–185.

Paarsch, Harry J., and Bruce Shearer, “Piece Rates, Fixed Wages, and Incentive Effects: Statistical Evidence from Payroll Records,” *International Economic Review,* Department of Economics, University of Pennsylvania and Osaka

University Institute of Social and Economic Research Association, Vol. 41, No. 1, 2000, pp. 59–92.

Podgursky, Michael, Ryan Monroe, and Donald Watson, "The Academic Quality of Public School Teachers: An Analysis of Entry and Exit Behavior," *Economics of Education Review,* Vol. 23, 2004, pp. 507–518.

Porter, Andrew C., Peter Youngs, and Allan Odden, "Advances in Teacher Assessments and Their Uses," in Virginia Richardson, ed., *Handbook of Research on Teaching*, Washington, DC: American Educational Research Association, 2001.

Prendergast, Candice, "The Provision of Incentives in Firms," *Journal of Economic Literature*, Vol. 37, No. 1, March 1999, pp. 7–63.

Rivkin, Steven G., Eric A. Hanushek, and John F. Kain, *Teachers, Schools and Academic Achievement,* Working Paper No. 6691, Cambridge, MA: National Bureau of Economic Research, 1998.

Santibañez, Lucrecia, Georges Vernez, and Paula Razquin, *Education in Mexico: Challenges and Opportunities,* Santa Monica, CA: RAND Corporation, DB-480-HF, 2005. Online only: http://www.rand.org/pubs/documented_briefings/DB480/.

Santizo, Claudia, *Implementing Reform in the Education Sector in Mexico: The Role of Policy Networks*, thesis, United Kingdom: University of Birmingham, 2002.

Schacter, John, and Yeow-Meng Thum, "Paying for High and Low-Quality Teaching," *Economics of Education Review*, Vol. 23, 2004, pp. 411–430.

Schmelkes, Sylvia, "Teacher Evaluation Mechanisms and Student Achievement: The Case of Carrera Magisterial in Mexico," paper presented at the annual meeting of the Comparative and International Education Society, Washington, DC, March 14–17, 2001.

Secretaría de Educación Pública (SEP), *Acuerdo Nacional para la Modernización de la Educación Básica*, México D.F., México: SEP, 1992.

———, *Plan de Estudios, Licenciatura en Educación Primaria*, México D.F., México: SEP, 1997.

———, *Plan de Estudios, Licenciatura en Educación Secundaria,* México, D.F., México: SEP, 1999.

Shepard, Lorrie A., and Katherine C. Doughtery, *Effects of High-Stakes Testing on Instruction*, paper presented at the annual meeting of the American Educational Research Association, Chicago, IL, 1991 (ERIC ED 337 468).

Stockard, Jean, and Michael Lehman, "Influences on the Satisfaction and Retention of 1st-Year Teachers: The Importance of Effective School Management," *Educational Administration Quarterly,* Vol. 40, No. 5, 2004, pp. 742–771.

Tokman, Andrea P., *Evaluation of the P900 Program: A Targeted Education Program for Underperforming Schools,* Working Paper No. 170, Santiago: Banco Central de Chile, 2002.

Trochim, William M., *The Research Methods Knowledge Base, 2nd Edition,* online at: http://trochim.human.cornell.edu/kb/index.htm (version current as of October 20, 2006).

Tyler, Elenes, and Nora Esperanza, "Carrera o Barrera Magisterial? Un Estudio Preliminar del Impacto en los Docentes de Primaria," *Pensamiento Universitario*, Vol. 86, 1997, pp. 82–99.

UNESCO-OREALC, *Laboratorio Latinoamericano de Evaluación de la Calidad de la Educación: Primer Estudio Internacional Comparativo Sobre Lenguaje, Matemática y Factores Asociados para Alumnos de Tercer y Cuarto Grado de Educación Básica,* UNESCO: Santiago, Chile, 2002.

Urquiola, Miguel, and Emiliana Vegas, "Arbitrary Variation in Teacher Salaries. An Analysis of Teacher Pay in Bolivia," in Emiliana Vegas, ed., *Incentives to Improve Teaching: Lessons from Latin America,* Washington, DC: The World Bank, 2005.

Vegas, Emiliana, and Ilana Umansky, "Improving Teaching and Learning Through Effective Incentives. Lessons from Latin America," in Emiliana Vegas, ed., *Incentives to Improve Teaching: Lessons from Latin America,* Washington, DC: The World Bank, 2005.

Webb, Norman L., *Criteria for Alignment of Expectations and Assessments in Mathematics and Science Education,* Washington, DC: National Institute for Science Education, University of Wisconsin-Madison, Council of Chief State School Officers, 1997.

Wise, Arthur E., Linda Darling-Hammond, Harriet Tyson-Bernstein, and Milbrey Wallin McLaughlin, *Teacher Evaluation: A Study of Effective Practices,* Santa Monica, CA: RAND Corporation, R-3139-NIE, 1984.

APPENDIX A

Table A.1
Means of CM Point Scores and Raw Test Scores by Selected Categories: Primary Teachers (*Etapas* 8–12)

Scores	Global	Education	Seniority	Student Achievement	Teacher Knowledge	Peer Review	Prof. Develop. (Natl.)	Prof. Develop. (State)	Raw Student Test Score	Raw Teacher Test Score
Total	62.5	8.5	7.2	11.9	17.3	9.3	8.2	4.9	40.9	48.7
Gender										
Female	62.5	8.5	7.3	12.0	17.1	9.3	8.2	2.7	41.7	48.3
Male	62.5	8.6	7.0	11.8	17.6	9.2	8.2	2.4	39.7	49.3
Seniority Categories										
0–4 years	57.8	9.0	2.5	11.9	17.1	9.1	8.1	2.6	38.2	47.8
5–9 years	59.4	8.9	3.7	12.0	17.3	9.2	8.1	2.6	39.4	48.4
10–14 years	60.4	8.5	5.4	11.9	17.2	9.2	8.1	2.7	40.7	48.6
15–19 years	62.8	8.4	7.0	12.0	17.6	9.3	8.2	2.8	41.1	49.6
20–24 years	64.1	8.4	8.5	12.0	17.4	9.3	8.2	2.6	40.9	48.7
Over 25 years	64.4	8.4	9.9	11.7	16.9	9.3	8.1	2.1	41.7	47.5
Education										
No degree	59.8	8.0	8.6	11.1	15.8	9.0	8.1	2.1	41.0	44.9
Pasante	60.8	8.0	7.5	11.6	16.9	9.3	8.1	2.4	40.7	47.9
Normal Básica	61.1	8.0	7.8	11.7	16.7	9.2	8.1	2.3	41.0	47.1
Normal Licenciatura	63.1	9.0	5.8	12.3	18.0	9.3	8.2	2.9	40.5	50.3
Normal Superior	64.9	9.0	7.9	12.2	18.0	9.3	8.2	2.7	41.2	50.6
Graduate	70.5	12.0	6.7	12.8	19.7	9.5	8.4	3.3	41.0	54.8
Strata										
Rural, low development	61.9	8.5	7.2	11.8	17.1	9.2	8.1	3.1	39.6	47.1
Rural, marginal development	59.0	8.5	6.5	10.5	16.4	9.0	8.1	3.0	37.0	45.2
Urban, low development	62.6	8.5	7.6	11.6	17.2	9.3	8.2	3.3	41.4	48.2
Urban, medium development	63.3	8.6	7.7	11.5	17.6	9.5	8.2	3.3	42.6	49.8
Urban, marginal development	61.6	8.5	7.3	11.5	17.1	9.1	8.1	3.2	39.7	47.1

Table A.2
Means of CM Point Scores and Raw Test Scores by Selected Categories: Secondary Teachers (*Etapas* 8-12)

Scores	Global	Education	Seniority	Student Achievement	Teacher Knowledge	Peer Review	Prof. Develop. (Natl.)	Prof. Develop (State)	Raw Student Test Score	Raw Teacher Test Score
Total	62.50	9.10	7.28	14.20	19.45	9.35	5.76	3.81	36.95	50.07
Gender										
Female	68.97	9.10	7.28	14.16	19.06	9.35	9.98	4.94	38.70	50.70
Male	68.73	9.08	7.23	14.13	18.94	9.20	10.26	4.91	35.37	49.49
Seniority Categories										
0-4 years	62.47	9.06	2.52	13.92	18.38	9.12	9.88	4.92	36.13	48.84
5-9 years	64.78	9.09	3.72	14.18	18.82	9.14	9.93	4.91	36.35	49.75
10-14 years	66.59	9.08	5.37	14.07	18.74	9.19	10.26	4.92	36.63	49.33
15-19 years	68.48	9.07	6.97	14.08	18.87	9.27	10.30	4.93	37.04	49.61
20-24 years	70.66	9.13	8.55	14.24	19.35	9.33	10.00	4.93	37.14	50.83
Over 25 years	71.64	9.08	9.92	14.18	19.17	9.34	10.03	4.93	37.35	50.81
Education										
No degree	67.86	8.00	8.38	12.89	17.67	9.15	11.45	4.92	37.34	46.14
Pasante	65.68	8.00	6.45	13.58	18.42	9.16	10.52	4.91	36.77	49.11
Normal Básica	68.94	8.00	8.10	12.34	17.10	9.33	11.96	4.95	37.33	46.53
Normal Licenciatura	69.62	9.00	8.40	13.91	18.52	9.19	10.59	4.94	37.67	49.45
Normal Superior	68.38	9.00	7.17	14.37	19.11	9.27	9.50	4.92	36.95	50.43
Graduate	74.57	12.05	7.47	14.82	20.80	9.44	8.95	4.94	37.03	54.41

Table A.3
Average Promotion and Incorporation Rates by Region (*Etapas* 8-12): Primary Teachers

Development Zone	**Not Incorporated (%)**	**Incorporated (%)**	**Promoted (%)**	**Total (%)**
Rural low development	94.22	0.9	4.87	100
Rural marginal development	95.73	0.71	3.56	100
Urban, low development	93.53	0.91	5.56	100
Urban, medium development	93.05	0.93	6.02	100
Urban, marginal development	94.32	0.74	4.94	100

APPENDIX B

Standards Related to Test Contents and Test Specification

AERA 3.2

The purpose of the test, definition of the domain, and the test specifications should be stated clearly so that judgments can be made about the appropriateness of the defined domain for the stated purpose(s) of the test and about the relation of items to the dimensions of the domain they are intended to represent.

AERA 3.3

The test specifications should be documented, along with their rationale and the process by which they were developed. The test specifications should define the content of the test, the proposed number of items, the item formats, the desired psychometric properties of the items, and the item and section arrangement. They should also specify the amount of time for testing, directions to the test takers, procedures to be used for test administration and scoring, and other relevant information.

AERA 3.5

When appropriate, relevant experts external to the testing program should review test specifications. The purpose of the review, the process by which the review is conducted, and the results of the review should be documented. The qualifications, relevant experiences, and demographic characteristics of expert judges should also be documented.

AERA 3.7

The procedures used to develop, review, and try out items and to select items from the item pool should be documented. If the items were classified into different categories or subtests according to the test specifications, the procedures used for the classification and the appropriateness and accuracy of the classification should be documented.

Standards Related to Test Construction

AERA 3.17

When previous research indicates that irrelevant variance could confound the domain definition underlying the test, then to the extent feasible, the test developer should investigate sources of irrelevant variance. Where possible, such sources of irrelevant variance should be removed or reduced by the test developer.

AERA 3.6

The type of items, the response formats, scoring procedures, and test administration procedures should be selected based on the purposes of the test, the domain to be measured, and the intended test takers. To the extent possible, test content should be chosen to ensure that intended inferences from test scores are equally valid for members of different groups of test takers. The test review process should include empirical analyses and, when appropriate, the use of expert judges to review items and response formats. The qualifications, relevant experiences, and demographic characteristics of expert judges should also be documented.

AERA 3.9

When a test developer evaluates the psychometric properties of items, the classical, or Item Response Theory (IRT) model used for evaluating the psychometric properties of items, should be documented. The sample used for estimating item properties should be described and should be of adequate size and diversity for the procedure. The process by which items are selected and the data used for item selection, such as item difficulty, item discrimination, and/or item information, should also be documented. When IRT is used to estimate item parameters in test development, the item response model, estimation procedures, and evidence of model fit should be documented.

AERA 7.3

When credible research reports that differential item functioning exists across age, gender, racial/ethnic, cultural, disability, and/or linguistic groups in the population of test takers in the content domain measured by the test, test developers should conduct appropriate studies when feasible. Such research should seek to detect and eliminate aspects of test design, content, and format that might bias test scores for particular groups.

ECIEE 4.2

Items will be written with basis on the specifications by academic personnel with the appropriate qualifications and training for the task.

ECIEE 4.3

Item contents and their alignment with the standards will be reviewed by personnel different from the item writers. (Related to AERA, APA and NCME, 1999, 3.5)

ECIEE 16.3

If a test includes items that do not meet standards of quality, or if the administration does not meet these standards, test users should be informed about the technical implications of these shortcomings. Such a situation should be rare and the problems addressed promptly so that they do not extend to more than one administration. If this is not possible, the test should not be operational and should be considered to be in piloting stages.

NRC (p. 76)

The development process should ensure balance and adequate coverage of relevant competencies. The development process should ensure that the level of processing required (cognitive relevance) of the candidates is adequate.

NRC (p.77)

The assessments should be field tested on an adequate sample that is representative of the intended candidates.

Standards Related to Test Reliability

AERA 1.1

A rationale should be presented for each recommended interpretation and use of test scores, together with a comprehensive summary of the evidence and theory bearing on the intended use or interpretation.

AERA 1.3

If validity for some common or likely interpretation has not been investigated, or if the interpretation is inconsistent with available evidence, that fact should be made clear and potential users should be cautioned about making unsupported interpretations.

AERA 2.1

For each total score, subscore, or combination of scores that is to be interpreted, estimates of relevant reliabilities and standard errors of measurement or test information functions should be reported.

AERA 2.2

The standard error of measurement, both overall and conditional (if relevant), should be reported both in raw score or original scale units and in units of each derived score recommended for use in test interpretation.

AERA 2.19

When average test scores for groups are used in program evaluations, the groups tested should generally be regarded as a sample from a larger population, even if all examinees available at the time of measurement are tested. In such cases the standard error of the group mean should be reported, as it reflects variability due to sampling of examinees as well as variability due to measurement error.

AERA 2.7

When subsets of items within a test are dictated by the test specifications and can be presumed to measure partially independent traits or abilities, reliability estimation procedures should recognize the multifactor character of the instrument.

AERA 4.3

If there is sound reason to believe that specific misrepresentations of a score scale are likely, test users should be explicitly forewarned.

AERA 4.1

Test documents should provide test users with clear explanations of the meaning and intended interpretation of derived score scales, as well as their limitations.

AERA 4.17

Testing programs that attempt to maintain a common scale over time should conduct periodic checks of the stability of the scale on which scores are reported.

AERA 4.2

The construction of scales used for reporting scores should be described clearly in test documents.

AERA 5.12

When group-level information is obtained by aggregating the results of partial tests taken by individuals, validity and reliability should be reported for the level of aggregation at which results are reported.

AERA 11.2

When a test is to be used for a purpose for which little or no documentation is available, the user of is responsible for obtaining evidence of the tests' validity and reliability for this purpose.

AERA 13.19

In educational settings, when average or summary scores for groups of students are reported, they should be supplemented with additional information about the sample size and shape or dispersion of score distributions.

AERA 15.7

When educational testing programs are mandated by school, district, state, or other authorities, the ways in which test results are intended to be used should be clearly described. It is the responsibility of those who mandate the use of tests to identify and monitor their impact and to minimize potential negative consequences. Consequences resulting from the uses of the test, both intended and unintended, should also be examined by the test user.

ECIEE 5.2

For each global, partial, or composite score, detailed information should be provided about reliability and standard error of measurement that allow the user to determine whether the level of precision is adequate for the intended use of the test. (Equivalent to AERA, APA and NCME, 1999, 2.1)

ECIEE 8.2

Scales used to report results of a test and the rationale for choosing such scales should be documented and published so as to ease interpretation of the results by test takers and the general public. These publications should make clear how scaled scores were created from raw scores. (Equivalent to AERA, APA and NCME, 1999, 4.1, 4.2)

ECIEE 6.1

Test developers will systematically oversee the different sources of evidence to assess the validity of a test, understanding validity as a unitary concept that involves content and construct analysis, and concurrent and predictive criteria. For each instrument, validity analyses should be started in the pilot phase and complemented with subsequent analysis periodically. The results should be published.

NRC (p.80)

At the time an assessment is first released, the development process should be clearly describe. Content-related evidence of validity should be presented along with any other empirical evidence of validity that exists; plans for collecting additional logical and empirical validity evidence should be provided and updated or modified as needed; results from these additional validation studies should be reported as soon as the data are available. In particular, the committee’s criteria should include the following: a comprehensive plan for gathering logical and empirical evidence for validation should specify the types of evidence that will be gathered (e.g., content-related evidence, data on the test’s relationships to other relevant measures of candidate knowledge and skill, and data on the extent to which the test distinguishes between minimally competent and incompetent candidates), priorities for the additional evidence needed, designs for data collection, the process for disseminating results, and a time line.

Standards Relating to Test Administration and Security

AERA 3.19

The directions for test administration should be presented with sufficient clarity and emphasis so that it is possible for others to replicate adequately the administration conditions under which the data on reliability and validity, and, where appropriate, norms were obtained.

AERA 5.1

Test administrators should follow carefully the standardized procedures for administration and scoring specified by the test developer unless the situation or a test taker's disability dictates that an exception should be made.

AERA 5.2

Modifications or disruptions of standardized test administration procedures or scoring should be documented.

AERA 5.6

Reasonable efforts should be made to assure the integrity of test scores by eliminating opportunities for test takers to attain scores by fraudulent means.

AERA 8.11

In educational testing programs and in licensing and certification applications, when it is deemed necessary to cancel or withhold a test taker's score because of possible testing irregularities, including suspected misconduct, the type of evidence and procedures to be used to investigate the irregularity should be explained to all test takers whose scores are directly affected by the decision. Test takers should be given a timely opportunity to provide evidence that the score should not be canceled or withheld. Evidence considered in deciding upon the final action should be made available to the test taker on request.

ECIEE 9.10

Printing of the tests and testing materials should be supervised to prevent errors, guarantee the security of the materials, and protect the confidentiality of the results. Leftover materials should be securely disposed.

ECIEE 9.11

Once printed, tests and materials will be kept in a secure place and remain under surveillance to guarantee that they (or their contents) will not be improperly distributed.

ECIEE 11.4

Staff responsible for test administration should receive sufficient training to carry out her designated tasks. Selection of such staff should decrease sources of bias, for example by not using a teacher to administer the test for her own students, or using individuals that are likely to have a particular interest in the results.

ECIEE 11.5

When users are in charge of administering the tests, clear instructions should be provided that emphasize the key aspects of administration and allow replication of the conditions of administration under which reliability and validity data for the test were obtained. (Equivalent to AERA, APA, and NCME, 1999, 3.19.)

ECIEE 11.6

Security measures should be in place to prevent information leaks during transportation and storage of the testing materials prior to and after administration. Such measures would ideally include secure transportation and storage, and sealed boxes, as well as double controls of the numbers of tests and monitoring during key steps in the distribution of the materials. It could also involve security committees, or even public notaries.

ECIEE 12.6

There should be one qualified person responsible for test application. This person should be the only one authorized to make decisions that can modify test application conditions in case of any unforeseen circumstances. This same individual should be informed of any suspicion of possible irregularities, so that annulment of any test takers results can take place.

ECIEE 13.8

In some cases it may be advisable to annul the results for one or more test takers because of irregularities or suspicion of fraud. Criteria and procedures should be in place for these types of decisions, of which the test takers should be informed. For high-stakes decisions (such as admission or licensure), every effort should be made to expedite the investigation and protect the interests of the test taker. The test taker should be notified the reason for suspicion and should have the right to offer relevant evidence in his or her favor. (Equivalent to AERA, APA and NCME, 1999, 8.10–8.13.)

Standards Relating to Test Information and Communication

AERA 5.10

When test score information is released to students, parents, legal representatives, teachers, clients, or the media, those responsible for testing programs should provide appropriate interpretations. The interpretations should describe in simple language what the test covers, what scores mean, the precision of the scores, common misinterpretations of test scores, and how scores will be used.

AERA 6.1

Test documents (e.g., test manuals, technical manuals, user's guides, and supplemental material) should be made available to prospective test users and other qualified persons at the time a test is published or released for use.

AERA 8.1

Any information about test content and purposes that is available to any test taker prior to testing should be available to all test takers. Important information should be available free of charge and in accessible formats.

AERA 11.15

Test users should be alert to potential misinterpretations of test scores and to possible unintended consequences of test use; users should take steps to minimize or avoid foreseeable misinterpretations and unintended negative consequences.

AERA 11.18

When test results are released to the public or to policymakers, those responsible for the release should provide and explain any supplemental information that will minimize possible misinterpretations of the data.

AERA 11.5

Those who have a legitimate interest in an assessment should be informed about the purposes of testing, how test will be administered, the factors considered in scoring examinee responses, how the scores are typically used, how long the records will be retained, and to whom and under what conditions the records may be released.

ECIEE 9.1

In addition to test administration materials (i.e., test forms, answer sheets) additional documentation for each test should include guides for test takers and test users, scoring guides, and forms to report results. (Related to AERA, APA and NCME, 1999, 8.1.)

ECIEE 9.2

Guides for test takers should provide the necessary information so they have a clear understanding of the characteristics and implications of the test. All relevant information should be made available to all test takers. (Related to AERA, APA and NCME, 1999, 8.1.)

ECIEE 14.3

Information provided to test users, decisionmakers, or the general public should include, in addition to descriptions of the purpose and characteristics of the test, details on what it can and cannot measure, the conclusions and decisions it can support, and any other information that helps prevent inappropriate interpretations of the results. (Related to AERA, APA, and NCME, 1999, 11.18.)

ECIEE 14.10

When test results are disseminated to the media, information should also be offered to help minimize the potential for incorrect interpretations. (Equivalent to AERA, APA and NCME, 1999, 5.10.)

Standards Relating to Test Documentation

AERA 3.1

Tests and testing programs should be developed on a sound scientific basis. Test developers and publishers should compile and document adequate evidence bearing on test development.

ECIEE 2.1

Test developers will have a technical manual with general specifications applicable to the tests they develop; when adequate, a separate manual will be available with specific technical information for each test. The manuals should be published or be available for review to any interested party.

ECIEE 2.2

The technical manual should be readily available to potential users before starting operational use of the test. (Equivalent to AERA, APA, and NCME, 1999, 6.1.)

ECIEE 2.3

Manuals should describe the theoretical foundation of each test in detail, as well as the psychometric models used in its development. The manual will indicate recommended uses for the tests and the support for these uses. They will include a summary of reliability and validity analysis (and others discussed below), with references to allow readers to locate sources and additional literature. When it is possible to anticipate inappropriate uses of the test, the manual should include the necessary warnings. (Equivalent to AERA, APA, and NCME, 1999, 6.3.)

Standards, Relating to Organizational Structure of the Tests and Testing Program

ECIEE 1.1

Test development organizations will maintain an organizational structure proportionate with the scale and importance of their activities and should include a minimum of elements and collegiate bodies adequate to ensure its adequate operation. In all cases, the organizational structure will include: (excerpt)

- An academic component with content specialists in the fields relevant for a test.

- A technical component, with specialists in test construction. If an organization develops more than one test this technical component can be a central unit common to all tests.
- A social component that includes users and representatives from other sectors like professionals, business leaders, teachers, and parents.
- A collegiate body (Technical Advisory Committee) occupying the highest rank in the testing program’s organizational structure. This body should include specialists from the three categories above and try to achieve balance in technical expertise.

APPENDIX C

METHODOLOGY AND SELECTED RESULTS

To explore the interrelationships between the program's measured criteria, we estimate multivariate regression models for the three measures of educational quality: (1) student test scores; (2) teacher test scores; and (3) peer review. The model for student test scores is as follows:

$$\mathrm{A}_{\mathrm{jt}} = \mathrm{Criteria}_{\mathrm{jt}}\beta_1 + \mathrm{X}_{\mathrm{jt}}\beta_2 + \mathrm{E}_{\mathrm{t}}\beta_3 + \varepsilon_{\mathrm{jt}} \qquad (1)$$

where the dependent variable, A_{jt}, is the average achievement of students of teacher j at time t, measured by the percentage of correct answers on the classroom test. The explanatory variables of interest (denoted by the row-vector criteria) are the measures capturing the remaining five CM promotion criteria: (1) percentage correct answers on the teacher performance test; (2) an indicator for whether the teacher took a state or national professional development (PD) course, and a continuous variable for teacher score on the professional development component; (3) teacher education dummies; (4) seniority in years; and (5) teacher score on the peer review component. β_1 is a column vector of parameters that capture the association between the promotion criteria variables and student achievement. We use standardized scores (subtract mean and divide by standard deviation) for student achievement, teacher performance, peer review and professional development, so that the coefficients on these variables (β_1) capture effect sizes.

In addition, we include other explanatory variables (denoted by the row-vector X) that are correlated with teacher scores and student achievement (gender, shift, subject, grade, development area, state). β_2 is a column vector of coefficients for these variables. We also include dummy variables for each evaluation year, or *Etapa* (denoted by row vector E) in all our regressions to control for any general time trends in student achievement. β_3 is a column vector that includes the coefficients for the *Etapa* dummies. In alternate models, we include teacher scores on the three sections of the teacher performance test instead of the overall teacher test score. This analysis allows us to determine the relative importance of each area of the teacher test on student achievement.

Estimation of the model in (1) using Ordinary Least Squares (OLS) may overstate or understate the relationship between the promotion criteria measures and student achievement due to unobserved confounding factors that remain in the error term. For example, teachers with high levels of ability or motivation are likely to perform better on all six promotion criteria. If factors like ability and motivation were not controlled for then the estimates of β_1 are likely to be biased upward.

In order to address this source of bias, we estimate teacher Fixed Effect (FE) models. Teacher FE models are akin to estimating the model in (1) including dummies for individual teachers. By doing so, the model uses a teacher as his control, and therefore eliminates all unobserved fixed factors that are correlated with student achievement and the factor scores. The coefficient estimates from a teacher FE model capture how changes in the explanatory variables for a teacher are associated with changes in his classroom test scores.

A second set of models uses the standardized teacher performance score of teacher *j* at time *t* (TS_{jt}) as the dependent variable.

$$\mathrm{TS}_{jt}=\mathrm{Criteria}_{jt}\alpha_1+\mathrm{X}_{jt}\alpha_2+\mathrm{E}_t\alpha_3+\upsilon_{jt} \qquad (2)$$

The explanatory variables of interest are the promotion criteria variables denoted by the vector criteria. These include: (a) whether or not the teacher took any PD course; (b) a continuous measure of the CM point scores for PD; (c) education dummies; and (d) seniority in years. The variable vectors X_{jt}, and E_t were defined earlier. α_1, α_2, and α_3 are the coefficient vectors corresponding to the three sets of variables in the model. We estimate (2) using OLS and teacher FE.

The final set of models is similar to the model in (2) but includes the standardized score on the peer review component as the dependent variable. The explanatory variables in these models include: (a) standardized score on teacher test; (b) standardized student achievement score; (c) whether or not the teacher took any PD course; (d) standardized PD score; (e) education dummies; (f) seniority in years; and (g) dummies for each *Etapa*.

While teacher FE models are a potentially powerful strategy to obtain unbiased estimates of the relationships between student achievement (or teacher performance) and other promotion criteria, there are some limitations of this approach that deserve mention. First, this approach only eliminates the bias that results from unobserved characteristics of a teacher that remain constant over time. For example, if teachers' motivation changes with time, a teacher FE approach may still yield biased estimates of the relationship between factor scores and student achievement. Second, this approach does not allow one to estimate the effect of fixed variables, such as gender or race, on student achievement. The teacher FE approach assumes that only factors that vary over time influence changes in the outcome of interest.

MISSING DATA AND PATTERNS OF TEACHER PARTICIPATION

To be included in the models discussed previously requires that observations have nonmissing data in all the variables included in the model. Unfortunately, a large proportion of the observations in the original CM sample are lost to missing data. In some instances missing data are due to the design of the program, and as such, would not introduce bias into our results as long as we restrict these results to apply to the population of *eligible* teachers under analysis. Other teachers, however, have missing data for reasons that could potentially be related to the outcome of interest and could therefore introduce significant bias into our estimates.

Table C.1 describes the nature of the data used for this analysis. In all the analysis conducted for this study, teacher positions are used as the unit of analysis (a teacher can hold two or more teaching positions (e.g., one position teaching in the morning shift and another position teaching in the afternoon shift) and is counted as two or more teachers; see chapter 3 for more on teaching positions). However, because some of the models include individual teacher fixed effects, Table C.1 also reports the number of individual teachers these positions correspond to.

Table C.1
Sample Used in Regression Analysis (*Etapas* 8-12)

Sample Description	**Number of Teachers (positions)**	**Number of Teachers (individuals)**
Primary Teacher Sample		
Total Sample	1,788,627	426,542
Sample for whom students were tested[a]	1,039,034	
Sample for whom student performance data was available	706,274	
Sample for student raw score regressions	648,833	269,740
Sample for teacher raw score regressions	1,215,905	379,500
Sample for peer review regressions	648,833	269,740
Secondary Teacher Sample		
Total Sample	402,165	108,143
Sample for whom students were tested[b]	307,650	
Sample for whom student performance data were available	203,730	
Sample for student raw score regressions	198,770	73,449
Sample for teacher raw score regressions	268,718	93,099
Sample for peer review regressions	198,770	73,449

[a] Among primary teachers, student testing was not done for positions in grades first and second, teachers in incomplete or multigrade schools, or teachers in rural boarding or indigenous schools.

[b] Among secondary teachers, student testing was not done for certain subjects, including technological education, physical education, and telesecondary teachers.

Actual sample sizes used to estimate our regression models are substantially lower than the original sample sizes in the CM database. This is in part because not all teachers' students are tested (see chapter 3). At the primary level, 70 percent of the data is lost due to teachers whose students are not tested. At the secondary level, this number is 45 percent. Of the teachers whose students were tested and still had missing student test score data, a further exploration of the data revealed that failure to take either the teacher test or professional development was the most likely cause behind the missing data. Of the primary teachers with missing student data, 87 percent had not taken the teacher test. This number was 92 percent for secondary teachers. As discussed in chapter 3, CM's guidelines state that students of teachers who do not take the teacher test are not tested (this prevents the program from spending resources on teachers who will receive a zero total score due to the missing teacher test).

The numbers in Table C.1 show that less than 10 percent of the observations at both the primary and secondary levels are lost due to missing data in all the regressions used in the analysis (see final regression sample vs. sample of teachers with test score data).

Table C.2 shows the comparison of scores and characteristics among teachers with missing and non-missing student test score data. Its results suggest minor differences in most variables, except whether teachers underwent professional development.

Unfortunately, we are unable to discern the direction of the bias these missing data could introduce in our analysis. It is possible that teachers who fail to take professional development or the teacher test are those who believe they would not do as well. This would introduce positive biases in some of our results, as the teachers in the regression sample would be comparatively "better" in their teacher and PD scores. However, it is also possible that teachers who participated in the past and obtained high scores on these measures do not plan to be evaluated in subsequent years, but nevertheless sign up for the program (and show up in the beginning sample) for fear of losing their current benefits.[63]

[63] Interviews with CM officials revealed that this was an increasing problem.

Table C.2
Comparison of Characteristics and Scores of Teachers with and without Student Test Score Data Whose Students Were Eligible to Take Student Tests

Sample Description	Teachers with Student Test Score Data	Teachers with Missing Student Test Score Data
Primary Teacher Sample		
Number of teachers	706,211	332,823
Teachers with *Normal Básica* (%) (basic 2- or 3-year teachers' college)	28	23
Teachers with *Normal Licenciatura* (%) (4-year teachers' college)	50	57
Average seniority (years)	18	19
Female teachers (%)	57	53
Average teacher test score (% correct answers)	48	46
Teacher took professional development (state or national) (%)	70	18
School operates the morning shift (%)	75	73
School operates the afternoon shift (%)	25	27
School is in rural marginal area (%)	15	14
School is in medium development urban area (%)	45	43
Secondary Teacher Sample		
Number of teachers	197,377	110,273
Teachers with *Normal Superior* (%)(4-year teacher's college for secondary teachers)	81	82
Average seniority (years)	18	18
Female teachers (%)	48	42
Average teacher test score (% correct answers)	51	50
Teacher took professional development (state or national) (%)	67	14
School operates the morning shift (%)	76	73
School operates the afternoon shift (%)	22	24
School is in rural marginal area (%)	n/a	n/a
School is in medium development urban area (%)	n/a	n/a

Table C.3 shows teacher participation by year. The numbers in this table refer to individual teachers and the number of times they are observed in our data with complete (i.e., nonmissing) information in all the variables analyzed.

Table C.3
Teacher Participation in *Etapas* 8-12:
Fixed Effects Models Regression Sample (Complete Cases Only)

No. of *Etapas* Teacher (Position) Participated In	Primary			Secondary		
	Teachers (Individual)	%	Cumulative %	Teachers (Individual)	%	Cumulative %
1	269,802	42.30	42.30	73,468	37.02	37.02
2	175,142	27.46	69.76	52,758	26.59	63.61
3	108,634	17.03	86.79	37,161	18.73	82.34
4	59,238	9.29	96.07	23,639	11.91	94.25
5	25,049	3.93	100.00	11,406	5.75	100.00
Total	637,865			198,432		

Table C.3 suggests that close to 40 percent of primary and secondary teachers are observed only once in our sample. More than 25 percent are observed twice, and close to 20 percent are observed three times. Fewer than 6 percent of primary and secondary teachers are observed in all five *Etapas* with nonmissing information in all the variables in the regression.

DESCRIPTIVE STATISTICS

Table C.4
Descriptive Statistics: Primary Teachers

Variable	Mean	Std. Dev. or % (for categorical variables)
Student test score (raw: % correct answers)	40.89	10.68
Teacher test score (raw: % correct answers)	48.26	12.00
Teacher test score (content: % correct answers)	49.35	15.72
Teacher test score (teaching methods: % correct answers)	47.75	13.67
Teacher test score (legal and other: % correct answers)	46.76	16.94
Supervisor/peer review rating (points)	9.22	1.12
Seniority (years)	18.57	7.27
Score on professional development test (points)	5.27	4.85
Teacher took professional development	0.71	0.45
No academic profile but with 15 years of teaching service	6,529	1.01
Technical studies completed at the mid-superior level	379	0.06
Advanced Teaching degree candidate or B.A. candidate	12,625	1.95
Basic teaching degree, or technical degree in teaching (*Normal Básica)*	325,303	50.14
Sixth semester in the B.A. in Indigenous Education (UPN)	851	0.13
B.A. degree related to specialty (special education)	2,601	0.4
Upper teaching degree for primary teachers (*Normal Licenciatura)*	185,524	28.59
Upper secondary teaching degree (*Normal Superior)* or B.A. degree related to subject	99,701	15.37
Master's or Ph.D. degree	15,320	2.36
Female	381,859	58.85
Male	266,974	41.15
Third grade	158,734	24.46
Fourth grade	157,425	24.26
Fifth grade	162,036	24.97
Sixth grade	170,278	26.24
Morning school	500,449	77.13
Afternoon school	147,090	22.67
Night school	1,294	0.2
Rural, low development area (RDB)	68,861	10.61
Rural, marginal development area (RM)	99,725	15.37
Urban, low development area (UDB)	112,800	17.39
Urban, medium development area (UDM)	289,313	44.59
Urban, marginal development area (UM)	78,134	12.04

Variable	**Mean**	**Std. Dev. or % (for categorical variables)**
Etapa 8	147,905	22.8
Etapa 9	135,568	20.89
Etapa 10	133,169	20.52
Etapa 11	120,539	18.58
Etapa 12	111,652	17.21
Observations (N)	648,833	

Table C.5
Descriptive Statistics: Secondary Teachers

Continuous Variables	**Mean**	**Std. Dev. or % (for categorical variables)**
Student test score (raw: % correct answers)	36.99	10.15
Teacher test score (raw: % correct nswers)	50.93	12.65
Teacher test score (content: % correct answers)	50.70	16.36
Teacher test score (teaching methods: % correct answers)	51.33	15.33
Teacher test score (legal and other: % correct answers)	51.28	18.54
Supervisor/peer review rating (points)	9.22	1.16
Seniority (years)	18.28	7.46
Score on professional development test (points)	5.32	5.12
Teacher took professional development	0.68	0.47
No academic profile but with 15 years of teaching service	3,794	1.91
Technical studies completed at the mid-superior level	607	0.31
Advanced teaching degree candidate or B.A. candidate	8,969	4.51
Basic teaching degree, or technical degree in teaching (*Normal Básica)*	600	0.3
Sixth semester in the B.A. in Indigenous Education (UPN)	34	0.02
B.A. degree related to specialty (special education)	854	0.43
Upper teaching degree for primary teachers (*Normal Licenciatura)*	2,270	1.14
Upper secondary teaching degree (*Normal Superior)*or B.A. degree related to subject	162,507	81.76
Master's or Ph.D. degree	19,135	9.63
Female teacher	95,195	47.89
Male teacher	103,575	52.11
Spanish	37,811	19.02
English	16,776	8.44
Math	35,056	17.64
Physics	11,087	5.58
Chemistry	15,139	7.62
Biology	20,626	10.38
Geography	11,974	6.02
History	25,454	12.81
Civics and ethics	18,588	9.35
Introduction to physics and chemistry	6,259	3.15
Morning school	151,616	76.28
Afternoon school	43,648	21.96
Night school	3,506	1.76

Continuous Variables	**Mean**	**Std. Dev. or % (for categorical variables)**
Etapa 8	49,812	25.06
Etapa 9	41,684	20.97
Etapa 10	40,568	20.41
Etapa 11	34,134	17.17
Etapa 12	32,572	16.39
Observations (N)	198,770	

Table C.6
Fixed Effects Regression Results: Primary Teachers

	Student Test Score (Model 1) Coef-ficient	t	Student Test Score (Model 2) Coef-ficient	t	Teacher Test Score Coef-ficient	t	Peer Review Rating Coef-ficient	t
Student test score (z-score)							0.0074	9.6
Teacher test: subject matter			-0.0003	-0.1				
Teacher test: methodology			-0.0150	-6.2				
Teacher test: legal			0.0139	6.5				
Teacher test score (z-score)	-0.0073	-2.4					0.0033	3.2
education_2	-0.0349	-0.4	-0.0359	-0.4	-0.0375	-1.1	0.0271	0.8
education_3	0.0205	0.6	0.0208	0.6	0.0049	0.4	0.0274	1.9
education_4	0.0506	1.6	0.0511	1.6	0.0057	0.5	0.0183	1.4
education_5	0.0692	1.3	0.0702	1.3	-0.0214	-0.9	0.0280	1.0
education_6	0.0669	1.4	0.0671	1.4	-0.0138	-0.7	0.0159	0.8
education_7	0.0448	1.4	0.0450	1.4	-0.0101	-0.8	0.0172	1.2
education_8	0.0348	1.1	0.0352	1.1	0.0128	1.0	0.0229	1.7
education_9	0.0449	1.1	0.0449	1.1	-0.0894	-5.6	0.0204	1.2
Seniority	0.0028	1.5	0.0028	1.5	0.0002	0.3	0.0023	1.9
Peer review (z-score)	0.0510	9.7	0.0508	9.6	0.0125	6.9		
PD score (z-score)	0.0239	9.0	0.0241	9.1	0.0190	15.6	0.0084	9.0
Teacher took PD	0.0423	8.6	0.0420	8.5	0.0546	24.7	0.0371	16.8
Afternoon shift	-0.1409	-16.8	-0.1414	-16.9	-0.0049	-1.3	-0.0126	-3.6
Evening shift	0.4696	4.8	0.4694	4.8	-0.0434	-1.5	-0.0157	-0.5
Female	0.0006	0.0	0.0019	0.1	0.0140	2.4	0.0013	0.2

Note: All models also included dummy variables for each state, *estrato*, *Etapa*, and grade.

Definition of Education variables:

1. No academic profile but with 15 years of teaching service (the last 10 at the same Level-Modality)
2. Technical studies completed at the mid-superior level
3. Advanced teaching degree candidate or B.A. candidate (of degree related to teaching subject) with at least 75 percent completed
4. Basic teaching degree, or technical degree in teaching (2-, 3-, and 4-year plans) (*Normal Básica)*
5. Sixth semester in the B.A. in Indigenous Education (UPN)
6. B.A. degree related to specialty (special education)
7. Upper teaching degree for primary teachers (*Normal Licenciatura*) or UPN
8. Upper teaching degree for secondary teachers (*Normal Superior*) or B.A. degree related to teaching subject
9. Master's or Ph.D. degree

Table C.7
Fixed Effects Regression Results: Secondary Teachers

	Student Test Score (Model 1) Coef-ficient	t	Student Test Score (Model 2) Coef-ficient	t	Teacher Test Score Coef-ficient	t	Peer Review Ratings Coef-ficient	t
Student test score (z-score)							0.0062	3.7
Teacher test: subject matter			-0.0017	-0.6				
Teacher test: methodology			-0.0245	-7.8				
Teacher test: legal			-0.0029	-1.1				
Teacher test score: (z-score)	-0.0141	-3.2					0.0030	1.4
education_2	-0.0210	-0.4	-0.0359	-0.4	0.0050	0.3	0.0319	1.0
education_3	-0.0286	-1.0	0.0208	-1.0	-0.0208	-1.3	0.0301	1.8
education_4	0.0170	0.3	0.0511	0.3	-0.0131	-0.7	-0.0149	-0.5
education_5	0.0021	0.0	0.0702	0.0	-0.0267	-0.2	0.0617	1.0
education_6	-0.0258	-0.6	0.0671	-0.6	-0.0096	-0.3	0.0217	0.7
education_7	-0.0559	-1.7	0.0450	-1.7	0.0004	0.0	0.0259	1.3
education_8	-0.0211	-0.9	0.0352	-0.9	0.0040	0.3	0.0368	2.2
education_9	-0.0639	-2.1	0.0449	-2.1	0.0342	1.8	0.0452	2.5
Seniority	-0.0013	-0.6	0.0028	-0.6	0.0032	1.7	0.0102	3.3
Peer review (z-score)	0.0280	3.7	0.0508	3.7	0.0112	2.4		
PD score (z-score)	0.0237	6.2	0.0241	6.3	0.0355	13.0	0.0117	7.2
Teacher took PD	0.0554	8.9	0.0420	9.1	0.0704	14.5	0.0415	12.7
Afternoon shift	-0.1288	-8.0	-0.1414	-8.0	0.0069	0.6	-0.0200	-2.1
Evening shift	-0.0521	-0.5	0.4694	-0.5	-0.0028	-0.1	0.1294	2.2
Female	0.0467	1.4	0.0019	1.4	0.0175	0.8	0.0170	0.7

Note: All models also included dummy variables for each state, *Etapa*, and subject.

Definition of Education variables:

1. No academic profile but with 15 years of teaching service (the last 10 at the same level-modality)
2. Technical studies completed at the mid-superior level
3. Advanced teaching degree candidate or B.A. candidate (of degree related to teaching subject) with at least 75 percent completed
4. Basic teaching degree, or technical degree in teaching (2-, 3-, and 4-year plans) (*Normal Básica)*
5. Sixth semester in the B.A. in Indigenous Education (UPN)
6. B.A. degree related to specialty (special education)
7. Upper teaching degree for primary teachers (*Normal Licenciatura*) or UPN
8. Upper teaching degree for secondary teachers (*Normal Superior*) or B.A. degree related to teaching subject
9. Master’s or Ph.D. degree

METHODOLOGY AND SELECTED RESULTS ON THE ANALYSIS IMPACT OF THE SALARY BONUS DURING THE YEAR OF INCORPORATION (SUPPORTING CHAPTER 6)

CM's national guidelines specify that teachers must receive a total score of at least 70 points to obtain incorporation or promotion. As described in chapter 3, of the total point score available to every teacher, up to 70 points are determined by the teacher's background characteristics, such as formal education, experience, and professional development courses, and by the teacher test score. Up to another 10 points are determined by peer review, but these ratings are generally high (the sample mean is 9.1). All of these factors are assessed in advance of the collection of the student test scores. The scoring procedure is well publicized via the distribution of materials to teachers, as well as national and state-level websites.[64] It is important to note that the minimum 70-point cutoff is a necessary but not sufficient condition for incorporation. States rank teachers according to their global point score and allocate their budgets for incorporation for teachers above that cutoff. However, it is possible that the budget is exhausted before all teachers scoring above 70 receive incorporation. We do not believe that this affects the assumptions in the regression-discontinuity approach because the cutoff is well publicized (and the final cutoff is not known by the teacher). Therefore, teachers will make their decisions (e.g., to exert more or less effort) based on their probability of reaching the minimum 70-point cutoff score.

It is worth noting that teachers do not have a lot of time between the time they take the professional development test (November), the teacher test (March), and the timing of the student test (June). It could be argued that this is not enough time to affect student test scores. However, it can also be argued that teachers, particularly those who have taken the test before or have talked to colleagues about it, can obtain some sense a priori of how they would score on the teacher test. This would give them an idea of how much effort they would need to exert in the classroom.

In this context, let us consider how the program might alter teachers' incentives to improve student achievement. The final assessment score (Final Points) can be expressed as *Finalpoints = f(X,A)*, where X is a vector of the teacher's background characteristics and A is the mean achievement of the

[64] See http://www.sep.gob.mx/wb2/sep/sep_617_carrera_magisterial.

teacher's students. Likewise, $A = a(X, Z, e)$ where Z is a vector of student background characteristics that determine achievement (e.g., parental schooling) and e is the chosen effort of each teacher. Additional effort is presumed to be costly for teachers.

A substantial portion of teachers' scores are determined prior to the collection of student test scores. Assume that teachers possess sufficient knowledge to calculate $Initialpoints = f(X,0)$, or the score they would receive when $A = 0$. From this calculation, it is apparent that some teachers face weak incentives to improve their students' achievement. If $Initialpoints \geq 70$, then teachers already fall above the promotion cutoff and have no additional incentive to exert costly effort over the course of the year. Similarly, if $Initialpoints < 50$, then teachers cannot be promoted, even if they obtain the full 20 points awarded for student test scores. Again, their incentives to exert additional effort to improve those test scores are weak.

If $50 \leq Initialpoints < 70$, teachers face stronger incentives to improve student test scores. More specifically, they are assumed to choose a level of effort, e*, such that $Finalpoints = f(X, a(X, Z, e^*)) = 70$. Such teachers would move from minimal effort to maximal effort. This forms the basis of the first empirical strategy, akin to a regression-discontinuity design.

As teachers cross the 50-point threshold, they face incentives to substantially increase their achievement and, hence, their effort. This would be evidenced by a break in the relationship between Initialpoints and test scores. As teachers' Initialpoints increases, however, the additional achievement and effort required to reach the promotion cutoff becomes progressively smaller. Eventually, close to 70, the achievement and effort required to reach the promotion cutoff is minimal and there will be little break, if any.

The magnitude of the break at 50 will likely vary across teachers, who surely recognize that some classroom achievement will result from their own background characteristics (the X's) and the characteristics of their students (the Z's). Thus, some teachers with $Initialpoints = 50$ will find that less effort is required to reach 70 (e.g., those with higher-SES and higher-achieving students). Others will find that substantial effort is required (e.g., those with lower-SES and lower-achieving students).

As an initial test, we will estimate:

$$Testscore_i = \beta_0 + \beta_1 Initialpoints_i + \beta_2 (50 \leq Initialpoints < 70)_i + \varepsilon_i \qquad (1)$$

where classroom test score points of the *i* teacher are a function of *Initialpoints*, a dummy variable indicating a value between 50 and 70, and an error term. We will also add controls for a limited number of observed teacher variables, as well as school fixed effects that control for unobserved school, teacher, and student variables that are constant within schools. As a sharper test of the presence of a break in classroom test scores around the discontinuity, we will further estimate:

$$Testscore_i = \beta_0 + \beta_1 Initialpoints_i + \beta_2 (Initialpoints \geq 50)_i + \varepsilon_i \qquad (2)$$

within subsamples of teachers whose values of *Initialpoints* fall within successively narrower bands around 50. Again, additional specifications will control for observed teacher variables and school fixed effects.

A plausible explanation for a discontinuous relationship between *Initialpoints* and test scores—gauged by β_2—would be a program effect. Yet, it is possible that a sharp break would not be observed, even in the presence of an effect for many individual teachers. Suppose that some teachers with *Initialpoint* = 50 derive substantial disutility from the effort required to obtain higher test scores, such that the expected award would not outweigh the disutility. A related possibility is that teachers have relative little time between the time they take the teacher test (which is also part of the *Initialpoint* score) until their students take their own test to induce any significant changes in achievement. These significant changes might require significant levels of extra effort on the part of the teacher and in some cases might be perceived as unattainable given the classroom context.

MEAN REVERSION ANALYSIS RESULTS

Table C.8
Mean Reversion Analysis Results

	Primary Teachers			**Secondary Teachers**		
	Student test score (1)	**Teacher test score (2)**	**Peer review rating (3)**	**Student test score (4)**	**Teacher test score (5)**	**Peer review rating (6)**
Level A	1.327**	2.024***	0.024	0.861***	1.857***	0.267***
	(0.539)	(0.549)	(0.102)	(0.255)	(0.192)	(0.057)
Level B	0.777	1.607***	-0.021	0.300	1.628***	0.156**
	(0.557)	(0.556)	(0.106)	(0.326)	(0.246)	(0.069)
Level C	0.065	0.698	-0.085	0.256	0.939***	0.068
	(0.577)	(0.567)	(0.110)	(0.376)	(0.296)	(0.084)
Level D	-0.255	-2.101***	-0.311**	-1.732**	-2.082***	-0.204
	(0.681)	(0.611)	(0.130)	(0.773)	(0.614)	(0.150)
N	427643	424001	571633	161076	215502	289983
R-squared	0.62	0.82	0.60	0.76	0.83	0.53
Teacher fixed effects?	Yes	Yes	Yes	Yes	Yes	Yes
Etapa dummies?	Yes	Yes	Yes	Yes	Yes	Yes

Note: Robust standard errors, clustered at the teacher level, are in parentheses. All Level coefficients are interpreted relative to the excluded category of *No incorporado*. *Etapa* coefficients not shown.
* significant at 10%; ** significant at 5%; *** significant at 1%.

NOTES ON THE DATA (CHAPTER 6, SECTION ON THE ANALYSIS IMPACT OF THE SALARY BONUS AFTER INCORPORATION OR PROMOTION)

The analysis in Chapter 6 on the impact of the salary bonus after incorporation or promotion requires including a variable indicating whether teachers are unincorporated or incorporated to one of CM's five levels. This is not straightforward because CM data are divided into two files: "Ci" and "Bi." The Ci files contain observations on every teacher who enrolled in CM in a given *Etapa*, but they do not contain each teacher's specific level (although the data do indicate which teachers are entirely unincorporated). The Bi files contain teacher's CM levels, but they only contain observations for teachers that are incorporated or promoted in a given *Etapa*.

To recover as many levels for teachers as possible, including those in the Ci data, we applied the following algorithm. First, we identified teachers in the Ci files that are not incorporated (and, hence, seeking incorporation to Level A). Second, we identified the level of some teachers in the Ci files, based upon their observed level in the Bi file (noting that this can only be accomplished for teachers in their year of promotion). Third, we recovered levels for additional teachers in the Ci files, based on the same teachers' levels in preceding or following *Etapas* (i.e., a teacher with Level A in *Etapa* 8 but promoted to Level B in *Etapa* 10 is assumed to be Level A in *Etapa* 9). Fourth, we recovered levels for some teachers that are already incorporated but not promoted in the *Etapas* 8–12 (hence, we cannot identify whether they are Level A, B, C, D, or E). To do so, we use data from *Etapas* 1 to 7 to identify the last observed level for such teachers.

ATTRITION ANALYSIS (SUPPORTING CHAPTER 6, SECTION ON THE ANALYSIS IMPACT OF THE SALARY BONUS AFTER INCORPORATION OR PROMOTION)

Tables C.9 and C.10 show descriptive statistics (means and standard deviations) on the differences in CM point scores as well as raw student and teacher test scores for primary and secondary teachers in *Etapas* 8 to 12. Unless otherwise noted, the differences are statistically significant at least at the 95 percent confidence level.

The results in Table C.9 suggest that, in most cases, nonreturning teachers had lower seniority, peer review ratings, and student test scores than returning teachers. They had higher education and teacher test score results. These findings are consistent with the attrition findings estimated using the full sample in the previous section. Results for secondary teachers are shown in Table C.10.

Table C.9
Differences in Means of Teacher Characteristics and Scores for Returning and Nonreturning Teachers: Primary Teachers

	***Etapa* 8**		***Etapa* 9**		***Etapa* 10**		***Etapa* 11**	
	Mean	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**
Nonreturning vs. returning teachers in nonpromoted group - Primary Teachers								
Teacher test score	n/s		n/s	0.121	-0.932	0.1008	0.2699	0.0978
Student test score	-2.827	0.245	-1.129	0.1554	-0.962	0.117	-1.463	0.1265
Peer review rating	-0.101	0.0078	-0.057	0.0066	-0.114	0.0051	-0.062	0.0052
Seniority	-0.699	0.0262	-0.695	0.0249	-0.533	0.0211	-0.53	0.0233
Education	0.1239	0.0085	0.1259	0.0082	0.0941	0.0071	0.0947	0.0081
Nonreturning vs. returning teachers in promoted group - Primary Teachers								
Teacher test score	n/s		0.897	0.2243	-0.914	0.2441	1.6998	0.2854
Student test score	-1.502	0.468	n/s		-1.065	0.3066	-1.176	0.3725
Peer review rating	-0.854	0.0391	-0.445	0.0309	-0.892	0.0487	-0.21	0.0171
Seniority	-0.766	0.0499	-0.29	0.0374	-0.336	0.0348	-0.334	0.0403
Education	-0.178	0.035	0.1089	0.019	0.054	0.0169	0.0739	0.0203

Table C.10
Differences in Means of Teacher Characteristics and Scores for Returning and Nonreturning Teachers: Secondary Teachers

	***Etapa* 8**		***Etapa* 9**		***Etapa* 10**		***Etapa* 11**	
	Mean	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**
Nonreturning vs. returning teachers in nonpromoted group - Secondary Teachers								
Teacher test score	n/s		-1.285	0.252	n/s		n/s	
Student test score	n/s		n/s		-0.372	0.282	-0.682	0.271
Peer review rating	-0.059	0.021	n/s		-0.073	0.015	-0.063	0.015
Seniority	n/s		0.123	0.045	n/s		n/s	
Education	0.100	0.024	0.049	0.021	n/s		0.150	0.0233
Nonreturning vs. returning teachers in promoted group - Secondary Teachers								
Teacher test score	n/s		n/s		-1.226	0.563	n/s	
Student test score	n/s		n/s		n/s		n/s	
Peer review rating	-0.912	0.096	-0.209	0.062	-0.929	0.1081	n/s	
Seniority	-0.301	0.135	0.203	0.097	n/s		0.302	0.112
Education	n/s		0.132	0.055	n/s		n/s	

The results in Table C.10 suggest that nonreturning and returning teachers are much more alike (i.e., their difference is not statistically different from zero) at the secondary level than at the primary level. Whenever differences were statistically significant, nonreturning secondary teachers had lower teacher and student test scores as well as peer-review ratings and higher education and seniority than returning secondary teachers.