INFORMES TÉCNICOS MIDE UC / TECHNICAL REPORTS MIDE UC
Centro de Medición MIDE UC / Measurement Center MIDE UC

IT 1002

Convergent Validity of the Chilean Standards-Based Teacher Evaluation System

MARÍA VERÓNICA SANTELICES AND SANDY TAUT

PONTIFICIA UNIVERSIDAD CATÓLICA DE CHILE
ESCUELA DE PSICOLOGÍA

Convergent Validity of the Chilean Standards-Based Teacher Evaluation System[1*]

MARÍA VERÓNICA SANTELICES[2] AND SANDY TAUT[3]

Abstract

This paper describes a convergent validity study of the mandatory, standards-based Chilean national teacher evaluation system (NTES). The study examined whether NTES identifies – and thereby rewards or punishes – the "right" teachers as high- or low-performing. We collected in-depth teaching performance data on a sample of 58 teachers who were evaluated by NTES as either "outstanding" (group 1) or "unsatisfactory" (group 2). The collected evidence included gains in student achievement scores, observation log data, expert ratings of a teaching materials binder, and teachers' scores on a subject and pedagogical knowledge test. The results support the validity of NTES' performance categorizations of the two extreme groups. The groups differed significantly on half of the performance indicators, and showed differences in the expected direction on the remaining indicators. We found especially strong and practically significant differences related to time on task during lessons, lesson structure, student behavior, and student evaluation materials. We also found significant correlations between our results and the sample scores on three out of four NTES instruments.

Keywords

Teacher performance, standards-based teacher evaluation system, validity, validation, Chile

---

[1] This is a preprint of an article submitted for consideration in the journal Educational Assessment 2010 (copyright Taylor & Francis); Assessment in Education is available online at http://www.tandf.co.uk/journals/titles/0969594X.asp. An earlier version of this paper was presented at the annual conference of the American Educational Research Assocation (AERA), April 9-13, 2007, in Chicago, USA.
* We wish to thank Carolina Thibaut, Ariela Simonsohn and Edgar Valencia for expert research assistance.
[2] Dr. M. Verónica Santelices, Facultad de Educación, Pontificia Universidad Católica de Chile. E-mail: vsanteli@uc.cl
[3] Dr. Sandy Taut, Coordinadora del Área de Investigación MIDE UC y profesora adjunta. Email: staut@uc.cl http://www.mideuc.cl/investigacion02_staut.php

Introduction

The purpose of this paper is to present a construct validity study of the Chilean national teacher evaluation system (NTES, or Docentemás). Since 2005 the evaluation is mandatory and is the basis for rewarding and sanctioning about 71,000 teachers working in the Chilean public education sector. This study is part of a validity research agenda of NTES developed by independent researchers at a university-based Measurement and Evaluation Center (Santelices, Taut, & Valencia, 2009; Taut, Santelices, Araya, & Manzi, in press). The evaluation distinguishes between "outstanding", "competent", "basic", and "unsatisfactory" performance. Performance standards guiding the evaluation have been defined, officially endorsed, published and widely disseminated as the "Marco Para la Buena Enseñanza [Guidelines for Good Teaching]" (Ministry of Education, 2004). The result of the evaluation has high-stakes consequences for individual teachers: outstanding and competent teachers are eligible for an increase in salary, unsatisfactory teachers are subject to mandatory professional development, and – if repeatedly evaluated "unsatisfactory" – loss of employment

The extent to which the instruments used by the NTES collect evidence that accurately reflects pedagogical effectiveness is a matter of importance at individual, municipal, and national levels. The NTES results are not only used to make important decision about teacher careers at the individual level. The information also informs local personnel decisions as well as broad national discussions on how to improve learning outcomes, reform school and classroom practices, and modify teacher education and licensing. Our study may bring legitimacy to the information provided by the NTES and to the decisions made based on that information.

The paper is also highly relevant for the U.S. educational context because there are school districts that recently started implementing standards-based teacher assessment and

incentive systems, for example, Cincinnati, Coventry and Washoe County (Milanowski, 2002; Heneman III, Milanowski, Kimball & Odden, 2006). As this is a relatively recent development in the U.S., not a lot of studies have been published about the validity of these evaluation systems (Heneman III, Milanowski, Kimball & Odden, 2006). Our paper presents an example of a validity study conducted on a teacher evaluation system that is similar in some of its basic characteristics to these U.S. examples. The study can thus serve to inform validation efforts in these contexts. In addition, it can inform the current discussion regarding the U.S. Department of Education's "Race to the Top" fund, which encourages the design of high-quality teacher and principal evaluation systems, defining teacher effectiveness as based on input from multiple measures, with students' achievement growth being a significant factor (U.S. Department of Education, 2009).

The paper first introduces the contextual background related to teacher evaluation in Chile and describes the national teacher evaluation system. We then discuss the literature on teacher evaluation in general, and on its validity in particular. The next chapters present the research questions, methods and findings of the construct validity study. Finally, we draw conclusions and suggest further research.

Contextual background

The Chilean educational system is decentralized and consists of three types of schools: municipal (public), private subsidized and private non-subsidized. In 2001, there were approximately 10,800 schools working in the system, 58% of which were municipal schools, 32% private subsidized schools and 10% private non-subsidized schools (Ministry of Education, 2003). Municipalities administer municipal schools, while private stakeholders (either individuals or private institutions) manage both private subsidized and private non-subsidized

schools.

In 2001 Chile had roughly 125,600 classroom teachers, of which 55% worked in municipal schools. Teachers currently do not have to pass a teacher licensure exam that would allow them to start their teaching practice. In municipal schools, teacher wages are linked to a state minimum wage, seniority, bonuses for additional training, geographic placement, and managerial responsibility, as well as bonuses that are based on an accreditation of excellence to schools (Sistema Nacional de Evaluación de Desempeño Profesional, SNED), and an individual certification of excellence (Asignación de Excelencia Pedagógica, AEP).

The national teacher evaluation system (NTES) was introduced by the Ministry of Education in 2003, and since 2005 is mandatory for teachers in municipal schools nation-wide. For more details on the development of the teacher evaluation system and its characteristics see Avalos and Assael (2007) as well as Manzi, Preiss, Gonzalez, Flotts and Sun (2008).

The evaluation system's formative, non-punitive character has consistently been stressed in official discourse (see, for example, Ministry of Education, 2003). At the same time, however, the NTES is a mandatory, high-stakes evaluation system where those teachers who are found to be high-performing are eligible for an increase in salary, while low-performing teachers are subject to professional development, and – if evaluated "unsatisfactory" in three consecutive years – loss of employment.

Evaluation methods include 1) portfolio assessment comprising a written part and a videotaped lesson, 2) supervisor assessment, 3) peer interview, and 4) self-assessment. The portfolio asks the teachers to describe planning and evaluation materials for a specific, pre-defined set of lessons, as well as to reflect on their use in the classroom. One lesson (45 minutes) of each teacher is videotaped by an external contractor. Two supervisors (generally the director

of the school and the teacher in charge of the so-called Technical Pedagogical Unit) complete an evaluation questionnaire asking about professional qualities of the evaluated teacher. The peer interview is performed by another teacher (not from the same school, but teaching the same subject and grade level) based on a structured interview protocol containing questions about pedagogical knowledge and practice. Finally, the self-assessment is a questionnaire that asks the teacher to critically reflect on his or her professional performance. The local evaluation commission can modify (up and down) the teacher´s final category based on the consideration of contextual variables detailed by the teacher, the peer evaluator and/or the supervisor.

The evaluated teachers receive a descriptive report detailing their results for the different portfolio dimensions and evaluation instruments. The school principal and the head of the municipal education authority receive summarized reports. The NTES 2008 results show that the majority (63.9%) of evaluated teachers received the performance categorization of "competent", while 22.8% were evaluated as showing "basic" performance. Only 12.8% were evaluated as "outstanding", and a mere 1% were considered as "unsatisfactory". Similar distributions of results were obtained by the NTES between 2003 and 2008. In total, so far 97 teachers have had to leave the public teaching force due to consecutive unsatisfactory performance.

Literature review

One of the most authoritative sources with regard to validation research is the Standards for Educational and Psychological Research (Joint Committee on Standards for Educational and Psychological Testing of the American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999). These Standards define validity as "the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests" (p. 9). The Standards agree with Kane (2001),

Messick (1994) and others that validation must focus on the proposed *interpretation* of test (or assessment) scores. Our construct validity study follows this suggestion by investigating the validity of NTES' interpretation of its collected evidence as a final, high-stakes categorization of teacher performance. The Standards differentiate types of validity, including content, construct and consequential validity; construct validity is further broken down into concurrent, convergent and discriminant validity. The Standards also identify the following sources of validity evidence: (a) test/instrument content, (b) response processes, (c) internal structure, (d) relations to other variables, and (e) consequences of testing. In this study we examined NTES' convergent validity by examining the relationship between teacher evaluation scores and their performance on different measures and variables related to teaching performance. We used *multiple* methods that are different from those currently used by NTES to collect data on participants' teaching performance.

Relevant literature also informed our choice of the most appropriate methods to validly and reliably assess teacher performance in our study. The literature recommends evaluating teacher performance by combining evidence gathered using a number of different methods (see Peterson, 2000; Joint Committee on Standards for Educational Evaluation, 1988). For each method, specific studies exist investigating their validity and reliability, for example, classroom observations were found to produce unreliable results if using limited time samples (Shavelson & Dempsey-Atwood, 1976; Shavelson, Webb & Burstein, 1986) and the paper-and-pencil assessment of teachers' subject-matter knowledge fell short of showing whether teachers were able to apply the knowledge in classroom situations (Shulman, 1987). Many recently developed performance-based teacher assessment systems in the U.S. (e.g., INTASC, Praxis) include a combination of the following data sources: (a) some collection of teaching materials related to

planning, instruction, student assessment, as well as including actual student work, (b) direct or video-taped observation of classroom performance, (c) teacher reflection on these types of evidence, and (d) an assessment of the teachers' subject-matter (and/or pedagogical) knowledge (see Porter, Youngs & Odden, 2001). Our construct validity study follows these examples, thus diverging somewhat from the evaluation methods used in the NTES itself.

The use of student achievement data, or learning gains of students, in the evaluation of teachers has been a controversial topic for decades (Millman, 1997). While there seems to be sufficient evidence that a teacher's classroom performance is one of the most important determinants of student learning (e.g., Nye, Konstantopoulos & Hedges, 2004; Wenglinsky, 2000), disparate views exist on

(1) whether a clear link between teacher performance and student achievement can be empirically established, considering the multiple factors at play in determining student achievement (e.g., Shavelson, Webb & Burstein, 2001; Darling-Hammond, 1999),
(2) how to measure and analyze student learning gains so that they can validly and reliably reflect differences in teacher performance (e.g., Lissitz, 2005; Kupermintz, 2003; McCaffrey, Lockwood, Koretz & Hamilton, 2003), and
(3) whether results of such value-added analyses should be included in teacher evaluation systems (Gordon, Kane & Staiger, 2006; Odden, 2004; Glass, 2004; Braun, 2005; Wright, Horn & Sanders, 1997).

While we are aware of the debates on the topic, we decided to include, as *one additional variable,* the assessment of student learning gains in our construct validity study, on the one hand because there is a vivid political debate also in Chile about using student achievement as an indicator in teacher evaluation, and on the other hand because we would have yet richer data that

would complement our teacher-based data collection methods.

Finally, we reviewed recent examples of validity studies of teacher evaluation, licensure and certification programs. For example, we closely studied the design of a construct and consequential validity study of the National Board for Professional Teaching Standards accreditation process (Bond, Smith, Baker & Hattie, 2000), and other NPBTS related validity studies (e.g., Goldhaber & Anthony, 2004; Pool, Ellett, Schiavone & Carey-Lewis, 2001). Pecheone and Chung (2006) examined different aspect of the validity of The Performance Assessment for California Teachers (PACT), while Le and Buddin (2006) provide a general overview of validity evidence for California teacher licensure exams. Wilson and Hallam (2006) recently presented a study using student achievement test scores as external validity evidence for indicators of teacher quality. All these studies, among others, informed parts of our design, instrument development, and analysis.

Research questions

The purpose of this study was to contribute to the comprehensive research program related to the validity of the Chilean national teacher evaluation system (NTES). This convergent validity study aimed at validating the final, high-stakes categorization of teachers, and we first focused on the "outstanding" and "unsatisfactory" performance categories. The relationship between the NTES score and the different measures used in our study provides information about the validity of the final NTES category: higher correlations indicate higher levels of validity and lower correlations indicate lower levels of validity. In addition the differences we find between the performance of unsatisfactory versus outstanding teachers on our measures will indicate whether NTES validly distinguishes especially high-performing from especially low-performing teachers.

The following research questions motivated the design of the study:

1. Based on the results of our study, do teachers found to be highest-performing by the NTES differ meaningfully in their teaching performance from teachers found to be lowest-performing by the NTES?
2. What is the correlation between the performance of participating teachers in the NTES evaluation of 2005 and in the validity study of 2006? In particular,
   (a) how do overall 2005 and 2006 categorizations correlate?
   (b) how do 2005 portfolio scores (written part and video analysis part) and 2006 observational and materials binder scores correlate?
   (c) how do other scores from 2005 evaluation instruments and 2006 validity study instruments correlate?

Methods

Below we briefly describe the sampling criteria, recruitment process, study design, and data analyses approaches we used in the study.

Sampling and recruiting of study participants

The universe for this study consisted of municipal teachers who had been evaluated by NTES in 2005. For our study we decided to focus only on elementary school teachers teaching 1st to 4th grade as they represent the largest subgroup of teachers evaluated in 2005. Among these, we recruited those teachers who taught either Mathematics or Language in the Metropolitan region in 2006. Finally, since the study aimed at validating those performance categories that are most consequential for teachers, we sampled only teachers who had received an either “outstanding” or “unsatisfactory” evaluation result. We strove to include 50% “outstanding” and 50% “unsatisfactory” teachers.

When we recruited the 58 teachers for our study, we asked them to take part in a "study on pedagogical practices" with one particular class of students throughout the whole year. For teachers in grades 2 and 3, we also restricted their participation to the subject area for which we had standardized student achievement tests available (Mathematics). We also asked the teachers to not reveal their NTES evaluation result to anyone working in the study.

The study offered two incentives to those teachers who would complete the entire study by December 2006: (a) a monetary incentive of $100,000 pesos (approx. US$180); and (b) a collection of teaching materials (copies) obtained from participating teachers as part of the study, presented in a binder and ordered by grade level and subject area.

Study design

The teachers we recruited for the study had to commit to completing the following activities throughout the course of the 2006 school year:

- Allow the administration of a curriculum-based student achievement test at the beginning and at the end of the school year;
- Allow the administration of a student background questionnaire at the beginning of the year;
- Allow for an observer to visit the classroom three times during the school year and, each time, audio-tape a 1.5 hour lesson and take notes (classes were not video-taped)[4];
- Collect planning, instructional and student evaluation materials (including actual

[4] Although the specific date of observation varied by teacher depending on the month of recruitment and the specific school and teacher calendar of activities, there was at least one month between each of the observations. In most cases the first observation took place between May and June, the second one between June and August and the third one between September and November.

student work) pertaining to a 2-week curricular unit in a structured binder; the binder that we provided also included two questionnaires on the teaching context and the teacher's reflections;

- Participate in a subject-related knowledge assessment and complete an end-of-year teacher questionnaire.

The study's goal was to collect as much in-depth data on actual classroom performance of the participating teachers as possible. Teacher performance was conceptualized based on the same set of teaching standards (MBE) as underlying the NTES, with one exception: student achievement was added as an indicator.

The research team either developed data collection instruments, or adapted existing instruments to the context of the study. Details on the data collection instruments can be found in Appendix 1.

Data analysis procedures

The following analyses were conducted: frequencies by items and groups of items for the overall sample, by grade level and subject matter; t-tests for mean equality between "outstanding" and "unsatisfactory" teachers; effect size calculations for significant mean differences; multilevel modeling in order to estimate the effect of teacher quality on students' standardized test performance; correlations between teacher performance as shown by our study's instruments and the NTES' instruments; reliability and internal consistency of our instruments and sub-scales.

Limitations of the study

There are several limitations related to this study. One concern is that the study took place the year following the NTES evaluation of those teachers who participated in the study.

This is problematic especially for the group of teachers who showed unsatisfactory performance. NTES is supposed to trigger change in these teachers by mandating to take professional development courses, and to be reevaluated the following year.

In addition, the paper focuses on the performance of outstanding and unsatisfactory teachers and does not study the two middle categories of NTES: basic and competent. Although these two categories amount to a large portion of teachers (approximately 75% of teachers in 2008), we have decided to focus on the extremes as they carry the most important consequences for teachers, namely the loss of employment and the opportunity to get a salary bonus. Furthermore, it is legitimate in early validation work to focus on the differences between the extreme performance categories first, before moving on to the more subtle distinctions between adjacent categories. We expect to investigate the differences between outstanding and competent teachers and between basic and unsatisfactory teachers in a future study.

Another threat to the generalizability of the study's findings is the self-selection of our sample of teachers who could have been motivated, at least in part, by the monetary incentive.

Finally, there are a number of issues related to the study's instruments and measurement procedures. The instruments were not pre-validated and their reliability coefficients were unknown. However we pilot tested all instruments used in the study and implemented several quality assurance measures such as scoring training and 13% of double observations. Further instrument validation, however, could have been possible in a study with a larger sample of teachers.

For the majority of participants the same research assistant performed all three classroom observations. A concern about this arrangement was that he or she would form an opinion of the teacher during the first observation and look to confirm this opinion during the second and third

observations. This concern applies mainly to the post-observation questionnaire.

Results

Sample descriptives

Participating teachers. The final sample on which all data analyses are based is N=58, (26 unsatisfactory and 32 outstanding). These teachers pertained to 22 different municipalities and 51 different schools. Nineteen of them participated with their language class, while 39 took part with their mathematics class. We had 15 first grades (7 in math class and 8 in language class), 12 second grades (10 in math class and 2 in language class), 16 third grades (all in math class) and 15 fourth grades participating (6 in math class and 9 on language class).

In terms of their performance on all four NTES instruments taken together (the portfolio, the supervisor assessment, the peer evaluation and the self-evaluation), it is interesting to note that of the 28 teachers assessed as "outstanding", N=27 received NTES portfolio scores of "competent" and N=1 received a portfolio score of "basic." As it is generally the case in NTES, our participants' final categorization as "outstanding" was due to their performance on the NTES' self-evaluation, peer interview, and supervisor assessment. With regard to the 25 study participants evaluated by NTES´ instruments as showing "unsatisfactory" performance, N=22 received portfolio scores of "unsatisfactory", N=1 received a portfolio score of "basic" and N=2 did not submit the portfolio. The local evaluation commission modified the final standing of the five remaining teachers included in the sample: 4 were raised from "competent" (based on the performance on the instruments) to "outstanding" and 1 was demoted from "basic" to "unsatisfactory."

Students. A total of 1,044 students participated both in the pre- and post-tests. Of the 531 students for whom we had information on their mother's education level, 30% had primary

education (complete or incomplete), while about 55% had secondary education (complete or incomplete). The remaining 15% completed a few years of technical post-secondary education.

Schools. Municipal schools in Chile are classified by the Ministry of Education into five socio-economic categories based on an index of social vulnerability of the school, mean family education and income. While none of the teachers in our sample came from schools pertaining to the extreme socio-economic categories, over half of the sample came from schools classified as mid-low. The distribution of teachers assessed as "outstanding" and "unsatisfactory" was fairly homogeneous in the medium and mid-low socioeconomic categories. However, in the schools serving students from mid-high socioeconomic background we were able to recruit only "outstanding" teachers.

Descriptive analyses and group comparisons

Teachers' standardized subject and pedagogical knowledge test. Overall the teachers' performance on the standardized subject and pedagogical knowledge test was poor. We think this may be in part explained by the fact that the test used is part of a program designed to certify teachers´ pedagogical excellence. Although it is based on the same standards used by NTES, both programs have very different goals. The proportion of correct responses (number of questions answered correctly over total number of questions) was 38% in Math and 37% in Language. We found statistically significant differences between "outstanding" and "unsatisfactory" teachers (see Table 1 for details); effect sizes (see separate section below) show moderate practical significance of these statistical differences.

Classroom observation. The 25 items from the *post-observation questionnaire* (see Appendix 1) were clustered into five factors based on a factor analysis we performed, considering 1) whether the items referred to lesson structure, 2) whether teachers used especially

stimulating instructional practices, 3) whether they made content-related or language-use mistakes, 4) whether students showed adequate behavior, and 5) whether the teacher flexibly adapted his or her teaching to the needs of the students. Since the indicators did not show a trend upward or downward over the course of the three observations, the t-test of mean equality was conducted using the mean score from the three observations. Results from this instrument comparing the practice of "outstanding" and "unsatisfactory" teachers on these five dimensions can be found in Table 1. Lesson structure and student behavior produced highly significant differences between these two groups of teachers, whereas stimulating instructional practices also accounted for significant differences. We found no significant differences regarding teacher conceptual or language errors (in general, very few errors were recorded) and teacher adaptation to the needs of the students.

Data from the *observation log* shows that "outstanding" teachers spent slightly (but non-significantly) more time on activities that are directly related to subject matter content and have a significantly larger proportion of students engaged in learning activities (on-task). For example, the proportion of time during which more than 95% of the students were on-task is significantly larger for "outstanding" than for "unsatisfactory" teachers, while the proportion of time during which less than 75% of the class was on-task is significantly larger for the "unsatisfactory" teachers (see Table 1 for details).

Teaching materials binder. Experts double-rated the different teaching materials binder sections using a set of pre-defined indicators as well as a holistic assessment. The scoring criteria were explicated in a detailed scoring rubric (see examples of the items and the rubric in Appendix 1). While some of the indicators were binary (scored as 0 if the attribute was missing or as 1 if it was present), most of them were scored on four-category scales as "unsatisfactory,"

"basic," "competent" or "outstanding," as was the holistic assessment. The instructional materials section was the one in which teachers did best as measured both by the indicators and by the experts' holistic assessment; they did worst in the reflection on their own practice section. It is important to keep in mind that the submission of materials in the binder was not mandatory, as we wanted the binder to reflect as closely as possible the teachers' actual daily practice (Manzi, Preiss, González, Flotts, & Sun, 2008). The section in which an important number of teachers did not submit materials was the one related to their interaction with parents and peers.

The results from the binder show that "outstanding" teachers tended to present higher quality instructional materials as holistically assessed by our experts, do a significantly better job at designing student assessments and providing feedback based on students' results, and are more reflective of their own practice than "unsatisfactory" teachers. Traditionally, student assessment design has been the area of weakest teacher performance at the national level, as indicated by the NTES portfolio results. Our findings confirm that the ability to design and use good student evaluation instruments clearly distinguishes between high-performing and low-performing teachers. No statistically significant differences between the two groups were observed in terms of the quality of their planning materials, and regarding the evidence related to their interaction with parents and peers (see Table 1 for details).

[Table 1 about here.]

Summary of the statistical significance of group comparisons. Table 1 shows t-tests and p-values for the group comparisons between "unsatisfactory" and "outstanding" teachers calculated based on the information collected by our study. Overall, 11 of the 22 statistical comparisons resulted statistically significant. This relates to 55% of all the comparisons based on the classroom observation data, 45% of all the comparisons based on the binder materials, and

50% of all the comparisons based on the teachers' subject matter test. The remaining comparisons all showed differences in the expected direction, that is, "outstanding" teachers outperforming "unsatisfactory" teachers.

Students' standardized tests. The students' standardized tests were administered in ten 2nd-grade classrooms, sixteen 3rd-grade classrooms, and fourteen 4th-grade classrooms. In both 3rd and 4th grades we used a different form for the pre- and post-test, but this was not possible for 2nd grade since there were too few items available. The pretest was administered to a total of 1,204 students and the post-test was administered to a total of 1,160 students. The proportion of correct responses (correct responses over total number of questions) for these students was 49.6% in the pre-test and 60.5% in the post-test. A subset of these students (n=1,044) was present at both the pre- and the post-test: 797 completed the Math tests and 247 completed the Language tests. The proportion of correct responses for the students that were present at both pre and post-test was 50.4% at pre-test and 60.7% at post-test, showing an average increase of 10.3% over the year.[5] Only students who participated in the pre- and post-testing were considered in the analyses that are presented in the remainder of this section, therefore there was no missing data for students.

When comparing student performance of "unsatisfactory" and "outstanding" teachers, we do not find statistically significant differences in pre-test scores, but we do observe statistically significant differences in post-test scores of these classrooms. That is, while we observe no statistically significant difference between the learning gains of "outstanding" and

[5] The students who took the pre-test and did not take the post-test showed poorer performance at pre-test than the students who took both tests (45% correct responses). The same trend was observed among the students who took the post-test but did not take the pre-test: they showed poorer performance than the students who took both tests (58.4% correct responses). The information reported here corresponds to the student-level.

"unsatisfactory" teachers' students when aggregating them at the teacher (classroom) level, we do find statistically significant differences in the expected direction when comparing learning gains of all students who were taught by "outstanding" teachers with the learning gains of students who who were taught by "unsatisfactory" teachers.

The results from the hierarchical linear modeling analysis on the importance of teacher quality on student achievement as measured by our standardized tests are not conclusive as they vary depending on the model used (see Table 2 for details).

All four models estimated were 2-level models, included achievement at the beginning of the year as a covariate and random effects in the level-1 intercept and slope coefficient. Interaction effects between teacher quality and initial achievement were tested and found to be not statistically significant. Therefore only the models without the interaction term are reported in Table 2. Models 1 and 3 were unconditional models.

[Table 2 about here.]

Following Raudenbush and Bryk (2002), the models were estimated using group-mean centering because of the important variation observed in the mean of the level-1 covariates by teacher (level-2 unit).[6] The teacher quality dummy variable was coded as follows: 1 = "outstanding" NTES performance category, 0 = "unsatisfactory" NTES performance category. Because of the limited sample size the analyses did not differentiate among grade levels or subject areas. Sample size did not allow us to include students´ or schools´ socioeconomic characteristics.

Model 2

[6] The mean achievement at the beginning of the year ranges from 15.2% to 84.2% of questions correct depending on the teacher and the test.

$$Achievement_{ij}^{2} = \beta_{0j} + \beta_{1j} * Achievement^{1}_{ij} + r_{ij}$$
$$\beta_{0j} = G_{00} + G_{1j} * TeacherQuality + u_{0j}$$
$$\beta_{1j} = G_{10} + u_{1j}$$

In Model 2 $Achievement_{ij}^{2}$ is the dependent variable and refers to the score of student i in classroom j at post-testing, $Achievement^{1}_{ij}$ is an independent variable and refers to the score of student i in classroom j at pre-testing, and r_{ij} is the random error component for student i in classroom j. In this model β_{0j} is the intercept term. The regression slope coefficient, β_{1j}, represents the effect of previous achievement on students´ observed achievement. This first equation of model 2 is referred to as the student-level (level-1) model since the observational units are students and each student`s outcome is represented as a function of his or her previous achievement. Controlling for student`s previous achievement is a standard practice when dealing with student learning and aims to identify the learning portion for which the particular teacher may be partially responsible for and control for the phenomenon known as "regression to the mean".

One of the goals of the analysis is to explain the average achievement of students (β_{0j}) and its relationship with teacher quality, which we defined as the teacher`s NTES final category (1 = "outstanding", 0 = "unsatisfactory"). This relationship is shown in the second equation of model 2 where G_{1j} gives the effect of teacher j on the average achievement of students (β_{0j}) and u_{0j} refers to the random error at the classroom level. The effect of previous achievement on students´ observed achievement (β_{1j}) is modeled as a random variable with error u_{1j}. The three error terms are assumed to have a normal distribution with mean of zero and variance τ_{rr}, $\tau_{u0,}$ τ_{u1} respectively.

In Model 4 the dependent variable was changed to represents Achievement Growth (i.e., the difference between scores at post- and pre-testing). All other variables of the model are the same as in model 2. In this case, controlling for previous achievement aims to control for the phenomenon called "regression to the mean" often observed in student assessment.

Model 4

$$AchievementGrowth_{ij}^{2} = \beta_{0j} + \beta_{1j} * Achievement^{1}_{ij} + r_{ij}$$
$$\beta_{0j} = G_{00} + G_{1j} * TeacherQuality + u_{0j}$$
$$\beta_{1j} = G_{10} + u_{1j}$$

While teacher quality as measured by the NTES is not statistically significant when using the difference in achievement (gain/growth/learning) as the level-1 outcome variable (model 4), previous achievement is statistically significant when considering the achievement at the end of the year as the level-1 outcome variable (p value of 0.01).[7] Although the level-2 variance component shows a reduction as a consequence of the introduction of teacher quality into the model both when comparing model 2 to the unconditional model (model 1) and model 4 to the unconditional model (model 3), the difference in explained variance between the two models is not substantial in terms of size. The overall deviance suggests that a significant proportion of the overall variance is not explained by the variables currently included in the model. However, we observe much smaller level-2 variance components for the gain score models (Models 3 and 4) as compared to Models 1 and 2, which indicates that between-teacher differences are much smaller in terms of students' gain scores than in terms of students' post-test scores.

[7] The fact that teacher quality is statistically significant is especially important if we consider the size of the sample: 39 teachers and 1,044 students.

The fixed effects estimated in Models 2 and 4 (see Table 2) can be interpreted as follows: In Model 2, the average mean pre-test achievement for all classrooms is 49.37 percent correct answers (G_{00}) (this is the class-level mean); the average difference in post-test achievement between students of "outstanding" and "unsatisfactory" teachers is 18.11 percent correct answers (G_{1j}), while controlling for students' pre-test achievement. In Model 4, the average mean gain achieved by all classrooms in the sample is 7.91 percent correct answers (G_{00}); the average difference in gain achieved by students of "outstanding" and "unsatisfactory" teachers is 3.97 percent correct answers (G_{1j}), while controlling for students' pre-test achievement.

Effect sizes. We calculated effect sizes (ES) for the t-tests of mean difference that were statistically significant. Considering an ES of 0.5 as one of medium or moderate practical significance and an ES of 0.8 as one of crucial importance (Hojat & Xu, 2004), we observe that all our effect sizes are either medium or large in size and have moderate to crucial practical importance (see Table 3). In terms of sources of information, we found that the indicators and sub-scales of the classroom observations show the strongest effects regarding differences between "outstanding" and "unsatisfactory" teachers, followed closely by the materials binder sections on evaluation design, student performance and reflection on own practice. The effect sizes related to the standardized teacher knowledge test indicate that the difference between "outstanding" and "unsatisfactory" teachers is only of moderate importance.

[Table 3 about here.]

Correlations between validity study and NTES instruments

This section presents the correlation between the performance of the teachers as shown by our study's instruments and their performance as measured by NTES' instruments.

Specifically, we were interested in looking at which of the four NTES instruments (self-evaluation, supervisor assessment, peer evaluation and portfolio) is/are more strongly correlated with our study's instruments. Also, the overall category based on all four NTES instruments was considered. Table 4 shows the largest (>0.3) of the statistically significant correlations.

From the data collected through the classroom post-observation questionnaire, correlations are particularly strong between lesson structure as well as appropriate student behavior and the NTES portfolio scores. NTES' final category correlates significantly with lesson structure, appropriate student behavior, and the proportion of time in which 95% of students are on-task. Lower, but nevertheless statistically significant correlations were also observed with especially stimulating instructional practices of the teachers.

The indicators of the teaching materials binder that most strongly correlate with the NTES instruments are those based on student evaluation design, student performance and instructional materials. The correlation between the performance on NTES' instruments and students' and teachers' standardized test performance are somewhat weaker and smaller than those observed in the observation data and binder assessment, but nevertheless important. It is interesting to note that the learning gains of students on our standardized tests do not correlate with any of the NTES instruments, but performance on the post-test correlates significantly with the peer evaluation and the supervisor assessment. Teacher performance on both the multiple-choice and open-ended items of the subject and pedagogical knowledge test correlates significantly with the NTES portfolio score.

[Table 4 about here.]

Internal consistency and reliability of instruments

Many instruments used in this study show satisfactory reliability indices: the post-

observation questionnaire, the sub-scales derived from this instrument, the teaching materials binder items considered as a whole, the 2nd grade and 3rd grade standardized student tests, and the multiple-choice section of the standardized teacher test. Some of the sections of the teaching materials binder show very poor reliability. Particularly low (or negative) are the reliability indices observed for the sections in which we found no statistically significant differences between "outstanding" and "unsatisfactory" teachers. We hypothesize that the lack of reliability is either due to the fact that these sections requested material that is easily available to teachers either through books or from the school curriculum specialist, or that the sections largely lacked materials (and that therefore the variance of scores in these sections is close to zero). The 4th grade math pre-test and the 4th grade language post-test also show low reliability. Therefore the t-tests of mean equality between "outstanding" and "unsatisfactory" teachers were conducted including all students in the sample first, and subsequently considering only students from 2nd and 3rd grades. The results were not significantly different.

We also calculated inter-observer reliability coefficients for 13% of the classroom observations, and for 100% of the binder ratings. For the classroom observations, we found acceptable inter-observer reliability for two out of five dimensions (0.73 for lesson structure and 0.80 for student behavior). For the remaining three dimensions (especially stimulating instruction, teacher errors and teacher flexibility) inter-observer reliability was low or negative. As for the binder ratings, we found very high inter-rater reliability for all sections except the planning materials section: between 0.82 and 0.96. Inter-rater agreement for the binder rating process was also calculated and found to be always significantly different from chance agreement, except for two rater pairs for the materials on communication with parents and peers.

Discussion

Results show that important differences between "outstanding" and "unsatisfactory" teachers are concentrated in their teaching practices related to lesson structure, student behavior, design of student evaluation materials, and their ability to ensure that all students are on-task most of the time. These differences are statistically significant and large in size. We observed differences of medium practical significance regarding teachers' subject knowledge, whether they used stimulating instructional practices, student performance as shown by the teachers' evaluation materials, and teachers' reflections about their own practice. No statistically significant differences between "outstanding" and "unsatisfactory" teachers were observed in the planning section of the binder, in the number and quality of teachers' communication with peers and parents as displayed in the binder, and in their students' learning gains during one school year.

A priori one could have expected that a larger proportion of the group comparisons would be statistically significant, considering that teachers in the sample come from the two extreme NTES categories. However, all the non-significant comparisons showed differences in the expected direction. In addition, the statistically significant comparisons were of strong practical significance (as shown by the medium and large effect sizes), and they seem to be concentrated in areas of more substantial pedagogical importance such as lesson structure, appropriate student behavior, students' time on-task, and design of student evaluation materials.

In terms of the correlations between our study's results and those of the NTES instruments, we find moderate correlations (between 0.3 and 0.6) for some indicators from the classroom observations, teaching materials binder, and teachers' standardized test performance. Although these indicators were most strongly correlated to the NTES' portfolio results,

classroom observations and binder assessments also correlated significantly with the NTES' supervisor and peer evaluations. The NTES' self-assessment results showed the lowest correlations with the performance on the validity study's instruments. Furthermore, it is interesting to note that there are important differences (as shown by statistical significance and effect size) between unsatisfactory and outstanding teachers in indicators that could have been most easily affected (improved) by motivation or professional development experienced by unsatisfactory teachers post NTES (e.g. lesson structure and student evaluation design).

The comparisons that were not statistically significant have alternative explanations that are worth considering. For example, the pedagogical knowledge section of the teachers' standardized test had only three questions, which could affect the power to detect statistically significant differences. The materials collected in two sections of the binder were non-informative: (i) the planning section in many cases was photocopied from a book or from materials given to the teacher by the school curriculum specialist, and (ii) the binders had only scarce materials regarding teachers' communication with peers and parents. The standardized tests used to assess student learning were not pre-validated and because of sample size limitations the multilevel analyses did not include more covariates. Multilevel analysis results are known to vary significantly from year to year, even in samples much larger than the one analyzed in this study. The student standardized tests varied significantly in terms of validity and reliability. In addition, the theoretical relationship between teacher performance, as defined by the "Marco Para la Buena Enseñanza [Guidelines for Good Teaching]" and student performance in standardized testing is far from clear (Ministry of Education, 2004).

To some degree the lack of important differences (as indicated by statistical significance and effect size) in some of the indicators was to be expected as these was an assessment system

implemented as a consequence of a political process in which the Teacher Union, local authorities and the Education Ministry participated. It is unlikely that a national teacher evaluation system would have included only indicators that would show important differences between teachers´ performance because the Teacher Union was aware that there are some indicators that are more challenging for teacher, for example, the subject matter and pedagogical knowledge test, or student achievement growth. Such a system would have been seriously resisted by the Teacher Union.

The study used indicators of teaching quality that were closely aligned with the standards framework underlying the NTES (Ministry of Education, 2004). These standards closely match those developed by Danielson (1996), which combine both the PRAXIS III and National Board for Professional Teaching Standards (NBPTS) efforts to develop standards for initial licensure on the one hand, and certification of excellence on the other hand. Danielson participated in these efforts at the Educational Testing Service and decided to provide a teaching framework for novice, midcareer and experienced teachers. This framework has since been used in U.S. school districts where standards-based teacher evaluation systems have been installed (see Odden & Kelley, 2002). Danielson's framework includes 22 teaching standards organized into four domains: planning and preparation, classroom environment, instruction, professional responsibilities. In Chile, a national consultation of teachers took place in which 80% of respondents validated the Chilean adaptation of these standards (Avalos & Assael, 2006). Since then they form the basis of the NTES.

Our study aimed at validating the NTES assessment itself, it did not aim at validating the teaching standards underlying it. Therefore, we incorporated indicators of teaching quality in our study that were aligned with the official teaching standards (see Tables 5 and 6 in Appendix 2).

The only exception is that we decided to test student achievement at the beginning and end of the school year to determine whether low-performing teachers as identified by NTES differ from high-performing teachers in the extent of student learning they generate. The student achievement aspect is excluded from the standards framework and from the NTES assessment. However, it is obvious, and NTES' stakeholders agree, that the final goal of the assessment is to contribute to improved student learning (Taut, Santelices, Araya & Manzi, in press). Establishing the link between teacher quality as diagnosed by NTES and student learning provides powerful evidence of the validity of this teacher performance assessment (see National Research Council, 2008).

The results from the analyses presented in this paper, although partial because they only refer to the two extreme NTES categories and to some aspects of validity, are positive. We found statistically significant differences in half of the indicators studied and better performance in the study´s instruments was correlated with better performance in NTES, particularly the portfolio. We interpret these results as convergent validity of the NTES classification.

Our analyses show that both outstanding and unsatisfactory teachers perform well in planning their classes and designing class materials. This may be due to the fact that books and other materials have been made available to schools in recent years and teachers have easy access to them. Teacher communication with peers and parents in a written format and disciplinary and pedagogical knowledge as measured by the subject matter test administered, on the other hand, do not seem frequent in any of the two groups studied. The study does not allow us to know whether effective teachers are actually communicating with peers and parents but through more informal ways. However, it does show that outstanding teachers create a more stimulating learning environment for their students, spend more time on task, present a clearer

lesson structure and design better student evaluation instruments. These are important elements to consider in the design of initial teacher education and professional development training courses.

Conclusions

The results presented above support the validity of NTES' final "outstanding" and "unsatisfactory" categories. We were able to observe practically and statistically significant differences between teachers classified as "outstanding" and "unsatisfactory" by NTES both when assessing their classes and when rating their teaching materials. There were also statistically significant differences between "outstanding" and "unsatisfactory" teachers in their performance on the standardized subject knowledge test, but the small mean difference between the two groups seemed of less pedagogical importance than those observed in the binder and classroom observation, as confirmed by the effect sizes. We found inconclusive evidence of differential student learning (as measured by standardized tests) associated to teacher performance categories.

NTES' validity is also supported by the correlational analyses we performed. We found moderate correlations especially for the NTES portfolio, and to a lesser extent for the supervisor and peer assessment. No relationship was found for the NTES self-assessment. This evidence should be considered in the future when discussing the weight each instrument should have in the final teacher performance category.

In a high-stakes teacher evaluation system like the one we find in Chilean public schools, research needs to build a sound validity argument that is constantly updated. Much validity research remains to be conducted. For example, we have initiated a consequential validity study whose purpose will be to study NTES' actual consequences for the primary intended users of the

evaluation results: evaluated teachers, school directors, and municipal education authorities. We start by (re-) establishing the theoretical underpinnings of the evaluation system: How is it *supposed to* improve teaching, and ultimately, student learning? Then we examine whether these intended consequences can in fact be observed among the main stakeholders. Besides tangible consequences such as salary increase, attendance at professional development courses, or non-renewal of work contracts, more intangible topics may include self-perception and work motivation of the teachers, school culture, and decision-making processes at the municipal level.

In the future, the validity of the contiguous categories of NTES should be explored, especially contrasting "unsatisfactory" and "basic" teacher performance.

The validity argument is built on an aggregation of evidence and we are in the process of amassing that evidence. This study is the first piece of information in a series of studies that will allows us to make a more definite judgment about the validity of a national assessment policy. Although not conclusive, our impression is that the results from this first approximation to the issue support the validity of Chile´s National Teacher Evaluation System.

References

Avalos, B., & Assael, J. (2007). Moving from Resistance to Agreement: The Case of the Chilean Teacher Performance Evaluation. *International Journal of Educational Research, 45*, 254-266.

Bond, L., Smith, T., Baker, W., & Hattie, J. (2000). *The Certification System of the National Board for Professional Teaching Standards: A Construct and Consequential Validity Study*. The University of North Carolina at Greensboro.

Borko, H. & Stecher, B. (2005). *Using Classroom Artifacts to Measure Instructional Practices in Middle School Mathematics: A Two-State Field Test.* CSE Report No. 662, National Center for Research on Evaluation, Standards, and Student Testing (CRESST), Graduate School of Education and Information Studies, University of California Los Angeles.

Braun, H. (2005). *Using Student Progress to Evaluate Teachers: A Primer on Value-Added Models*. Princeton, NJ: Educational Testing Service.

Danielson, C. (1996). Enhancing professional practice: A framework for teaching. Alexandria, VA: Association for Supervision and Curriculum Development.

Darling-Hammond, L. (1999). *Teacher Quality and Student Achievement: A Review of State Policy Evidence*. University of Washington: Center for the Study of Teaching and Policy.

Glass, G. (2004). *Teacher Evaluation.* Arizona State University.

Goldhaber, D., & Anthony, E. (2004). *Can Teacher Quality Be Effectively Assessed?* Washington, D.C.: The Urban Institute.

Gordon, R., Kane, T., & Staiger, D. (2006). *Identifying Effective Teachers Using Performance on the Job*. Washington, D.C.: Brookings Institution.

Heneman III, H., Milanowski, A., Kimball, S., & Odden, A. (2006). Standards-based Teacher

Evaluation as a Foundation for Knowledge- and Skill-based Pay. CPRE Policy Briefs RB-45, Consortium for Policy Research in Education, University of Pennsylvania, Graduate School of Education.

Hiebert, J., Gallimore, R., Garnier, H., Bogard Givvin, K., Hollingsworth, H., Jacobs, J. & others (2003). *Teaching Mathematics in Seven Countries: Results From the TIMSS 1999 Video Study.* Washington, D.C.: National Center for Education Statistics, U.S. Department of Education Institute of Education Sciences.

Hojat, M., & Xu, G. (2004). A Visitor's Guide to Effect Sizes. *Advances in Health Science Education, 9*, 241-249.

Joint Committee on Standards for Educational and Psychological Testing of the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education (1999). *Standards for educational and psychological testing.* Washington, D.C.: American Educational Research Association.

Joint Committee on Standards for Educational Evaluation. (1988). *The Personnel Evaluation Standards: How to Assess Systems for Evaluating Educators*. Newbury Park: Corwin.

Kane, M. T. (2001). Current Concerns in Validity Theory. *Journal of Educational Measurement, 38*(4), 319-342.

Kupermintz, H. (2003). Teacher Effects and Teacher Effectiveness: A Validity Investigation of the Tennessee Value Added Assessment System. *Educational Evaluation and Policy Analysis, 25*(3), 299-318.

Le, V., & Buddin, R. (2005). *Examining the Validity Evidence for California Teacher Licensure Exams*. Working Paper WR-334-EDU. Santa Monica: RAND.

Lissitz, R. (Ed.). (2005). *Value Added Models in Education: Theory and Applications.* Maple

Grove, MN: JAM Press.

Manzi, J., Preiss, D., Gonzalez, R., Flotts, P. & Sun, Y. (2008). Design and Implementation of a National Project of Teaching Assessment: The Chilean Experience. Paper presented at the annual meeting of the American Educational Research Association, March 24-28, 2008, New York City, USA.

McCaffrey, D., Lockwood, J. R., Koretz, D., & Hamilton, L. (2003). *Evaluating Value Added Models for Teacher Accountability*. Santa Monica: RAND.

Messick, S. (1994). The Interplay of Evidence and Consequences in the Validation of Performance Assessments. *Educational Researcher, 23*(2), 13-23.

Milanowski, A. (2002). The varieties of knowledge and skill-based pay design: A comparison of seven new pay systems for K-12 teachers. CPRE Research Report Series RR-050, Consortium for Policy Research in Education, University of Pennsylvania, Graduate School of Education.

Millman, J. (Ed.). (1997). *Grading Teachers, Grading Schools: Is Student Achievement a Valid Evaluation Measure?* Thousand Oaks: Corwin Press.

Ministry of Education (2004). *Marco Para La Buena Enseñanza* [Guidelines for Good Teaching]. Santiago: Ministerio de Educación.

Ministry of Education, Evaluation and Curriculum Unit (2004). *Chile y el aprendizaje de matemáticas y ciencias según TIMSS* [Chile and math and science achievement as measured by TIMSS]. Retrieved on November 25, 2006, from: http://www.simce.cl/doc/timms2003_imforme.zip

Ministry of Education (2003). *Attracting, Developing and Retaining Effective Teachers, OECD Activity, Country Background Report for Chile*. Prepared by the Ministry of Education,

Planning and Budget Division, Department of Research and Statistics. Retrieved on November 28, 2006, from: http://www.oecd.org/dataoecd/37/31/26742861.pdf

Myford, C. & Engelhard, G. (2001). Examining the Psychometric Quality of the National Board for Professional Teaching Standards Early Childhood/Generalist Assessment System. *Journal of Personnel Evaluation in Education 15*(4), 253-285.

Nye, B., Konstantopoulos, S., & Hedges, L. (2004). How Large Are Teacher Effects? *Educational Evaluation and Policy Analysis, 26*(3), 237-257.

Odden, A. (2004). Lessons Learned About Standards-Based Teacher Evaluation Systems. *Peabody Journal of Education, 79*(4), 126-137.

Pecheone, R. L., & Chung, R. (2006). Evidence in Teacher Education. The Performance Assessment for California Teachers (PACT). *Journal of Teacher Education, 57*(1). 1-15.

Peterson, K. (2000). *Teacher Evaluation* (2nd ed.). Thousand Oaks: Corwin Press.

Pool, J. E., Ellett, C., Schiavone, S., & Carey-Lewis, C. (2001). How Valid are the National Board of Professional Teaching Standards Assessments for Predicting the Quality of Actual Classroom Teaching and Learning? Results of Six Mini Case Studies. *Journal of Personnel Evaluation in Education, 15*(1), 31-48.

Porter, A. C., Youngs, P., & Odden, A. (2001). *Advances in Teacher Assessments and Their Uses*. In V. Richardson (Ed.), *Handbook of Research on Teaching*. Washington, D.C.: American Educational Research Association.

Raudenbush, S., & Bryk, A. (2002). *Hierarchical Linear Models. Applications and Data Analysis Methods*. London, New Delhi.: Sage Publications. Thousand Oaks.

Santelices, V., Taut, S., & Valencia, E. (2009). *Relacion entre los Resultados de la Evaluacion Docente y los Planes de Superacion Profesional: Estudio Descriptivo. [Relationship*

Between the National Teacher Evaluation System and Professional Development Courses: Descriptive Study]. Santiago: MIDE UC [Measurement and Evaluation Center, Catholic University].

Shavelson, R., & Dempsey-Atwood, N. (1976). Generalizability of Measures of Teaching Behavior. *Review of Educational Research, 46*(4), 553-611.

Shavelson, R., Webb, N., & Burstein, L. (1986). Measurement of Teaching. In M. Wittrock (Ed.), *Handbook of Research on Teaching* (3rd ed., pp. 569-598). New York: Macmillan.

Shavelson, R. J., Webb, N. M., & Burstein, L. (2001). Measurement of Teaching. In V. Richardson (Ed.), *Handbook of Research on Teaching*. Washington, D.C.: American Educational Research Association.

Shulman, L. S. (1987). Knowledge and Teaching: Foundations of the New Reform. *Harvard Educational Review, 57*(1), 1-22.

Taut, S., Santelices, V., Araya, C., & Manzi, J. (in press). The Theory Underlying a National Teacher Evaluation Program. Evaluation and Program Planning (2010). doi: 10.1016/i.evalprogplan.2010.01.002

U.S. Department of Education (2009, November). Overview information; Race to the Top Fund; Notice Inviting Applications for New Awards for Fiscal Year 2010. Retrieved on January 31, 2010 from: http://www2.ed.gov/programs/racetothetop/applicant.html

Wenglinsky, H. (2000). *How Teaching Matters. Bringing the Classroom Back Into Discussions of Teacher Quality*. Princeton, NJ: Educational Testing Service and Milken Foundation.

Wilson, M., & Hallam, P. (2006, April). Using student achievement test scores as evidence of external validity for indicators of teacher quality. *Paper presented at the Annual Conference of the American Educational Research Association,* San Francisco, CA.

Wolfe, E. & Gitomer, D. (2000). *The Influence of Changes in Assessment Design on the Psychometric Quality of Scores*. Princeton, NJ: Educational Testing Service.

Wright, P., Horn, S. P., & Sanders, W. (1997). Teacher and Classroom Context Effects in Student Achievement: Implications for Teacher Evaluation. *Journal of Personnel Evaluation in Education, 11*, 57-67.

Table 1

Statistical Significance of Group Comparisons

Indicator	Mean Outstanding Teachers	Std Dev	N	Mean Unsatisfactory Teachers	Std Dev	N	t-test of equal means: t	Degrees of freedom	Sig. (bilateral)
TEACHER KNOWLEDGE TEST									
Subject Matter Questions (45 multiple-choice questions)[1]	*18.09*	*5.57*	*32*	*14.85*	*5.78*	*26*	*2.17*	*56*	*0.034**
Pedagogical Knowledge Questions (3 open-ended questions)[1]	1.63	0.59	31	1.40	0.39	26	1.76	55	0.084
CLASSROOM OBSERVATION QUESTIONNAIRE									
Proportion of time in which activities were directly related to content	83.2	8.3	32	82.5	9.7	26	0.30	56	0.765
Proportion of time in which < 75% of students were on-task	11.7	10.6	32	26.0	18.6	26	*-3.70*	*56*	*0.001***
Proportion of time in which 75% to 95 % of students were on-task	44.0	14.3	32	45.7	16.6	26	-0.42	56	0.673
Proportion of time in which > 95 % of students were on-task	44.3	20.0	32	28.2	23.2	26	*2.84*	*56*	*0.006***
Lesson Structure	0.77	0.11	32	0.63	0.14	26	*4.21*	*56*	*0.000***
Especially Stimulating Instruction	0.28	0.20	32	0.17	0.17	26	*2.35*	*56*	*0.022**
Teacher Did not Commit Errors	0.96	0.07	32	0.93	0.13	26	1.24	56	0.220
Appropriate Student Behavior	0.84	0.11	32	0.69	0.17	26	*3.99*	*56*	*0.000***
Teacher Adaptation and Flexibility	0.06	0.12	24	0.08	0.13	22	-0.56	44	0.578
TEACHING MATERIALS BINDER									
Planning Section (continuous indicators)	0.57	0.11	32	0.56	0.11	25	0.377	55	0.707
Planning Section (holistic assessment)	0.71	0.18	32	0.69	0.23	25	0.226	55	0.822
Instructional Materials (continuous indicators)	0.97	0.05	32	0.95	0.11	26	0.941	56	0.351
Instructional Materials (holistic assessment)	0.87	0.14	32	0.80	0.16	26	*3.075*	*56*	*0.003***
Student Evaluation Design (continuous indicators)	0.82	0.10	31	0.73	0.04	24	*4.436*	*53*	*0.000***
Student Performance (continuous indicators)	0.57	0.21	31	0.46	0.19	23	*2.115*	*52*	*0.039**
Student Evaluation Design and Student Performance (holistic assessment)	0.59	0.21	31	0.41	0.14	23	*4.362*	*52*	*0.000***
Interaction with Parents and Peers (continuous indicators)	0.74	0.09	19	0.68	0.16	15	1.227	32	0.229
Interaction with Parents and Peers (holistic assessment)	0.70	0.20	19	0.65	0.26	15	1.085	32	0.286
Own-Practice Reflection Questionnaire (continuous indicators)	0.43	0.09	32	0.39	0.08	26	1.584	56	0.119
Own-Practice Reflection (holistic assessment)	0.55	0.15	32	0.44	0.15	26	*3.001*	*56*	*0.004***

* Indicates significance at alpha level of 5% and ** indicates significance at alpha level of 1%.

[1] The mean performance in the subject matter and pedagogical knowledge questions refer to the raw score. Each question answered correctly was scored as 1.

Table 2

Results from the Hierarchical Linear Model Analysis

Model	Level 1 Outcome Variable	Variables Included Level 1	Variables Included Level 2	Level 1 Variance Component	Level 2 Variance Component	Deviance	Fixed Effects		
							G_{00}	G_{1j} (Teacher Quality)	G_{10}
1	Achievement 2	Achievement 1	-	209.46	567.83	8734.65			
2	Achievement 2	Achievement 1	Teacher Category (Dummy Variable)	209.46	501.60	8720.31	49.37 (0.00)[1]	18.11 (0.01)	0.43 (0.00)
3	Achievement Growth	Achievement 1	-	209.28	54.52	8654.41			
4	Achievement Growth	Achievement 1	Teacher Category (Dummy Variable)	209.27	52.57	8646.39	7.91 (0.00)	3.97 (0.12)	-0.57 (0.00)

1 Numbers in parenthesis refer to significance level p.

Table 3

Effect Sizes (Cohen's d) for t-Tests that Showed Significant Mean Differences

Instrument	Indicator/sub-scale	Cohen's d
Teachers' content and pedagogical knowledge test	Multiple-choice items	0.58
Teachers' content and pedagogical knowledge test	Open-ended items	0.48
Observation log	Proportion of time in which less than 75% of students were on-task	0.98
Observation log	Proportion of time in which more than 95% of students were on-task	0.86
Post-observation questionnaire	Lesson structure	1.11
Post-observation questionnaire	Especially stimulating instruction	0.62
Post-observation questionnaire	Appropriate student behavior	1.05
Binder with teaching materials	Instructional materials (holistic assessment)	0.46
Binder with teaching materials	Student evaluation design (continuous indicators)	1.21
Binder with teaching materials	Student performance (continuous indicators)	0.58
Binder with teaching materials	Student evaluation design and student performance (holistic assessment)	0.95
Binder with teaching materials	Own practice reflection (holistic assessment)	0.75

Table 4

Correlations between Performance on the Study's Instruments and NTES' Instruments (Pearson Correlation and Significance)

	All Instruments	Portfolio	Peer Interview	Supervisor Assessment	Self-Assessment
Proportion of Time in which > 95% of Students on-task	0.39** (0.00)	0.329* (0.01)			
Lesson Structure	0.53** (0.00)	0.53** (0.00)	0.41** (0.00)	0.40** (0.00)	
Appropriate Student Behavior	0.49** (0.00)	0.49** (0.00)	0.37* (0.01)	0.32* (0.02)	
Especially Stimulating Instruction (Continuous Indicators)	0.30* (0.02)	0.32* (0.02)		0.36** (0.00)	0.36** (0.01)
Instructional Materials (Holistic Assessment)	0.52** (0.00)	0.50** (0.00)	0.36** (0.01)	0.33* (0.02)	
Student Evaluation Materials (Continuous Indicators)	0.56** (0.00)	0.58** (0.00)	0.36* (0.01)	0.43** (0.00)	
Student Performance (Continuous Indicators)	0.33* (0.02)	0.31* (0.03)			
Student Evaluation Materials and Student Performance (Holistic Assessment)	0.52** (0.00)	0.50** (0.00)	0.36** (0.01)	0.33* (0.02)	
Teachers' Test Scores (Subject Matter Knowledge)		0.41** (0.00)			
Teachers' Test Scores (Pedagogical Knowledge)		0.31* (0.02)			
Students' Standardized Post-Test Scores	0.37* (0.02)		0.41** (0.00)	0.41** (0.00)	

* Indicates significance at alpha level of 5% and ** indicates significance at alpha level of 1%.

Appendix 1: Data Collection Instruments

Student achievement tests were obtained for 2nd and 3rd grade Math, and 4th grade Math and Language. Third and fourth grade tests were constructed with items developed for SERCE, a comparative study implemented in Latin American by UNESCO's Regional Office for Latin America. These items had been piloted and validated as part of the SERCE study. The second grade Math test was developed by university researchers for evaluating student learning as part of an initiative by the Ministry of Education. All items were pilot-tested and different forms were used for pre- and post-testing. Pre-testing was done early in the school year, as soon as teachers agreed to participate and were able to schedule the test (usually May). Post-testing took place at least six months later, very close to the end of the school year, between November and December. Curriculum experts reviewed the items with regard to their reflection of the Chilean curriculum and age-appropriate item difficulty. The experts also helped build two parallel test forms for each test and developed scoring guides. There was no test for first grade students. We scored questions 0 (false) or 1 (correct) in most cases, used a scoring rubric and registered the total number of correct responses.

Class book copies were obtained to collect student background information (e.g., age, gender, mother's and father's education, special educational needs). We compared this information to the one collected through the student questionnaire and since these data were more complete and reliable we used it to describe the student sample.

The purpose of the *teacher observation log* was to create a structured account of teacher and student activities, as well as data to calculate the percentage of time spent on content, and the percentage of students on task. This instrument was inspired by the TIMSS 1999 video study (Hiebert et al., 2003) and developed by the researchers for this study. In order to obtain reliable

results we asked observers to record teachers´ and students´ behavior every three minutes over a total of 60 minutes, making it possible to record one entire lesson (45 min) from start to finish.

The goal of the *post-observation questionnaire* was to obtain an evaluative picture of the teacher's pedagogical skills. This instrument was based on expected teacher performance as outlined in the "Marco para la Buena Enseñanza" [Guidelines for good teaching] (Ministry of Education, 2004) and contained 25 items that were organized into five sub-scales: 1) lesson structure (Cronbach alpha of 0.75, 14 items), 2) appropriate student behavior (Cronbach alpha of 0.84, 2 items), 3) especially stimulating instruction (Cronbach alpha of 0.62, 3 items), 4) whether the teacher committed errors (Cronbach alpha of 0.66, 2 items), and 5) whether the teacher flexibly adapted to students' learning needs (Cronbach alpha of 0.46, 2 items). Ten of the 25 items were scored in a five-point Likert scale (never, seldom, sometimes, often, always, does not apply), one had a three-point Likert scale, thirteen were scored as Yes or No and there was one open-ended question. All Likert-scale and binary items were included in a factor analysis. We provide three example items below:

Sample items

Q13. The teacher associated explicitly at least one the class content to the students´ personal experiences. 0=No, 1=Yes.

Q21. There was a logic sequence in the order class activities were implemented. 0=Never, 1=Seldom, 2=Sometimes, 3=Often, 4=Always, 88=Does not apply.

Q22. The teacher used language properly (written and spoken)? 0=Never, 1=Seldom, 2=Sometimes, 3=Often, 4=Always, 88=Does not apply.

The observation instrument was pre-tested and the research assistants underwent training about its application. During actual observation the research assistants followed a standardized

protocol. Ten percent of all observations were conducted in pairs in order to determine the inter-observer reliability coefficient. In order to obtain reliable results we asked observers to record teachers´ and students´ behavior every three minutes over a total of 60 minutes, a period of time longer than those used by other researchers. How long is a regular class, 45 minutes? Would this also allow them to observe the whole class structure?

The *binder for collecting teaching materials* had six sections:

- planning materials;
- instructional materials used in the classroom;
- student evaluation materials;
- actual evidence of a student evaluation from the entire class;
- materials used to communicate with parents and/or collaborate with colleagues;
- two questionnaires: an "Identification Questionnaire" to be completed at the beginning of the materials collection process and a "Reflection Questionnaire" to be completed at the end of the materials collection process.

Completion of the questionnaires was obligatory, the rest of the binder was supposed to reflect actual teaching practice as closely as possible. We recommended teachers to include teaching materials from one instructional unit lasting approximately 2 weeks. NTES portfolio and this binder differ in that former asks teachers to work on a predefined set of learning objectives. NTES also asks teachers to complete all sections of the portfolio. In addition, the NTES portfolio includes a video-taped lesson. Teachers participating in the study had complete autonomy to choose the learning objectives covered by the binder. In our study we only collected written materials in the binder, since the observation protocol was designed to assess pedagogical practices in the classroom.

Below we show the scoring rubric of two of the items from the teaching material binder. Both items are part of the first section of the binder: planning materials. This section has a total of ten items. The first item is assessed using four points on a scale from 1 (unsatisfactory) to 4 (outstanding), and the second item is the holistic assessment of the planning materials section.

Presence or absence of the minimum elements for lesson planning. Although most of our items were assessed using four-levels of performance this particular item only has three since the task was considered a standard teaching practice.

4 Outstanding	3 Competent	2 Basic	1 Unsatisfactory
-------	Planning materials show all **six of the following six elements**: - Vertical Learning Objectives (OFV)*, - Expected learning outcomes, - Assessment indicators or instruments, - Learning contents, - Learning activities, - Horizontal Learning Objectives (OFT)*.	Planning materials show **five of the following six elements**: - Vertical Learning Objectives (OFV)*, - Expected learning outcomes, - Assessment indicators or instruments, - Learning contents, - Learning activities, - Horizontal Learning Objectives (OFT)*.	Planning materials show **four or less of the following six elements**: - Vertical Learning Objectives (OFV)*, - Expected learning outcomes, - Assessment indicators or instruments, - Learning contents, - Learning activities, - Horizontal Learning Objectives (OFT)*.

* The national curriculum refers to learning objectives which are specific to a particular grade level and discipline as "Vertical Learning Objectives" and to those across grade levels or subjects as "Horizontal Learning Objectives."

Holistic assessment of planning materials section.

Please provide a general assessment of this section. Use a number from 1 to 4, where 4 is the maximum score, to evaluate this section. Please consider the indicators assessed previously in this section. If there are other indicators that you consider relevant, please include them and make them explicit. Justify your judgment.

The *Reflection Questionnaire* had a total of 5 items scored from 1 (unsatisfactory) to 4 (outstanding) and the Cronbach Alpha was 0.60. Two example items are provided: (1) Please describe the relationship between the learning objectives and your students` performance, (2)

Please describe the relationship between your work environment and the learning outcomes of your students. as shown by the evidence of a student learning assessment you performend and included in this binder.

Subject and pedagogical knowledge test. We obtained permission to use the 2004 version of the standardized subject and pedagogical knowledge test for "Generalists" (teachers in elementary grades 1-4 teaching mathematics, language, and social science) used in a national program to certify teaching excellence. This test consisted of 45 multiple-choice items on subject matter knowledge, and three open-ended questions on pedagogical knowledge. The test items (both multiple-choice and open-ended questions) were scored using a scoring guide provided by the program. Each open-ended question had a four-point response scale ranging from "unsatisfactory" to "outstanding".

The *final teacher questionnaire* contained questions about professional background, professional development activities, job satisfaction, school context factors, and opportunity-to-learn data for the 2006 class. Teachers were asked to complete the questionnaire as part of the subject and pedagogical knowledge test.

Most of the teachers in our sample (41 out of 58) obtained their education degrees from a university while the rest studied in other kinds of higher education institutions such as professional institutes (their programs tend to be shorter and to have a more technical profile). On average, teachers had taught for 20 years and had been at the school where we observed them for about 11 years.

From the teacher questionnaire we also learned that 15 out of the 58 teachers attended some form of professional development in the subject matter of interest during the time of the study. In addition, 19 had participated in the accreditation process of teaching excellence, and 15

had obtained recognition from the Ministry of Education because of their school's performance. Most of the sample (41 teachers) declared a job satisfaction level that ranged between mid-high and very high.

Appendix 2: Alignment of NTES Portfolio (2005), Observation Log, Classroom Observation Questionnaire and Teaching Material Binder with "Marco para la Buena Enseñanza" (MBE) (2004) [Guidelines for Good Teaching]

Table 5

Alignment of Teaching Material Binder Sections, MBE Indicators and the NTES Portfolio Dimensions

Sections of "Teaching Material Binder"		Assessment criteria	MBE Criteria*	NTES portfolio dimensions
1. Planning & preparation	a)	Structure (logical sequence)	A.4	Planning of learning unit (1)
	b)	Clarity of language and instructions	C.3	
	c)	Alignment with curriculum	A.4	
	d)	Usefulness to guide instruction	-	
2. Evidence of teaching in class	a)	accuracy of contents	A.1	Planning of learning unit (1), Quality of classroom activities (2), Lesson structure (7)
	b)	accuracy of language	C.3	
	c)	motivating students	B.2/C.2	
	d)	engaging students actively in learning	B.2/C.2	
	e)	variety of instructional approaches	A.3	
	f)	alignment of activities with objectives of lesson	A.4	
	g)	alignment of activities with curriculum	A.4	
	h)	appropriateness of instructional approaches for age group	A.2	
	i)	appropriateness of activities for contents being taught	A.3/A.4	
	j)	resources used to support student learning	A.3	
3. Student evaluation instruments	a)	Alignment with learning objectives	A.4	Quality of student evaluation instrument (3)
	b)	Criteria for evaluating student learning are clear	C.1	
	c)	Feasibility of assessment administration	-	
	d)	Diversity of strategies	A.4	
	e)	Adequacy of strategies depending on contents to be assessed	A.4/C.6	
4. Student outcomes	a)	Learning gains visible?	A.3	-
	b)	Level of outcomes: higher order thinking skills? Provoking individual opinions? Creativity?	C.2/C.5	
5a. Exchange with colleagues	a)	Participation in meetings and professional development	D.2	-

	b)	Exchanges and discussion with colleagues about instruction and student learning		
5b. Communication with parents	a)	Provides information to parents about objectives and results of student learning	D.4	-
	b)	Motivates parents to get involved in education		
6. Reflection on practice	a)	Level of reflection: high-medium-low	D.1	Utilization of student evaluation results (4), Reflection on teaching practice (5)
	b)	Explanations related to student learning		
	c)	Teacher responsibility acknowledged?		
	d)	Professional development needs and interests identified		
	e)	Changes in practice		

*See footnote Table 6 for details on MBE criteria.

Table 6

Alignment of Observation Log, Classroom Observation Questionnaire, MBE and NTES Portfolio

Items	MBE criteria*	NTES portfolio indicators
OBSERVATION LOG		
Variety of instructional formats	A.3, A.4	-
Active involvement of students (their activities)	A.4	6.1
Time spent on learning subject matter	B.4, C.4	7.2
Percentage of class on task	-	-
CLASS OBSERVATION QUESTIONNAIRE		
CLASSROOM MANAGEMENT		
1. control over student behavior	B.3	6.3
2. respect	B.1	6.2,6.3
EQUAL OPPORTUNITIES FOR ALL STUDENTS		
3. Equality girls and boys	B.1, B.2	-
4. Equality active and passive students	B.2	-
USE OF TIME FOR OPPORTUNITIES TO LEARN		
5. efficient organization of lesson	B.4, C.4	7.1,7.2
6. clear orientation for tasks and activities	C.1	7.2
7. adaptation of class according to student needs	C.4	7.1
8. adaptation of class because of contextual changes		
9. adequate timing of classroom activities	C.4	7.2
CLASSROOM PEDAGOGY		
10. clearly communicated the objective(s)	C.1	-
11. motivated students for objective(s)	C.1	-
12. connection new contents to previous contents	C.2	8.1
13. connection new contents to students' personal experiences	C.2	-
14. addressed common errors or misconceptions	A.3	8.4
15. stimulated students' critical thinking	C.5	-
16. stimulated students' creativity	C.5	-

17. offered a summary of the main points	A.4, C.2	-
18. reinforced correct responses	-	8.4
19. corrected mistakes	C.5	8.4
20. students' interest level	B.2, C.1	-
21. classroom activities followed logical sequence	C.3	7.1
22. logical link between activities and objectives of the lesson	A.4	7.3
23. teacher used correct and precise language	C.3, C.5	-
24. teacher committed conceptual errors	C.3	-

*MBE criteria:

A. Preparation for teaching

A.1 Knows the contents of the subjects he/she teaches and the national curriculum.

A.2 Knows about his/her students' characteristics, knowledge and experiences.

A.3 Appropriately applies didactic approaches of the subjects he/she teaches.

A.4 Organizes learning objectives and contents in coherence with the curriculum and student characteristics.

A.5 Evaluation strategies are coherent with objectives, subject, curriculum and allow all students to demonstrate their learning.

B. Building a learning environment

B.1 Establishes a classroom climate characterized by acceptance, equity, trust, solidarity and respect.

B.2 Shows high expectations regarding all students' learning and development.

B.3 Establishes and maintains consistent norms regarding classroom functioning.

B.4 Develops a structured work environment, physical space and resources in service of student learning.

C. Teaching so that all students can learn

C.1 Communicates learning objectives clearly and precisely.

C.2 Teaching strategies are challenging, coherent and have meaning for the students.

C.3 Lesson contents are taught with conceptual rigor and students are able to understand them.

C.4 Makes optimal use of the available teaching and learning time.

C.5 Fosters cognitive development.

C.6 Monitors and evaluates students' learning processes and outcomes.

D. Professional responsibilities

D.1 Reflects systematically about his practice.

D.2 Builds professional relationships with his/her colleagues.

D.3 Takes responsibility in guiding his/her students.

D.4 Has respectful, collaborative relationships with parents.

D.5 Stays informed about new developments in his/her profession, the educational system and educational policies.