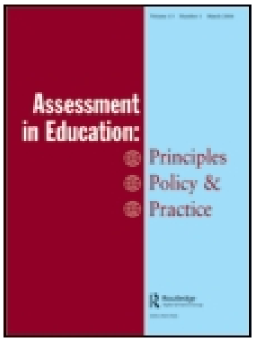

Assessment in Education: Principles, Policy & Practice

ISSN: 0969-594X (Print) 1465-329X (Online) Journal homepage: http://www.tandfonline.com/loi/caie20

Convergent validity evidence regarding the validity of the Chilean standards-based teacher evaluation system

Maria Verónica Santelices & Sandy Taut

To cite this article: Maria Verónica Santelices & Sandy Taut (2011) Convergent validity evidence regarding the validity of the Chilean standards-based teacher evaluation system, Assessment in Education: Principles, Policy & Practice, 18:1, 73-93, DOI: 10.1080/0969594X.2011.534948

To link to this article: https://doi.org/10.1080/0969594X.2011.534948

Published online: 25 Jan 2011.

Assessment in Education: Principles, Policy & Practice
Vol. 18, No. 1, February 2011, 73–93

Convergent validity evidence regarding the validity of the Chilean standards-based teacher evaluation system

Maria Verónica Santelices[a]* and Sandy Taut[b]

[a]*Facultad de Educación, Pontificia Universidad Católica de Chile, Santiago, Chile;* [b]*Escuela de Psicología, MIDE UC, Pontificia Universidad Católica de Chile, Santiago, Chile*

This paper describes convergent validity evidence regarding 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 categorisations 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 behaviour, 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

Introduction

The purpose of this paper is to present convergent validity evidence regarding 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 (Taut et al. 2010). 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,

*Corresponding author. Email: vsanteli@uc.cl

ISSN 0969-594X print/ISSN 1465-329X online

DOI: 10.1080/0969594X.2011.534948
http://www.informaworld.com

municipal, and national levels. The NTES results are not only used to make important decisions 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 relevant for the US 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 et al. 2006). As this is a relatively recent development in the US, not many studies have been published about the validity of these evaluation systems (Heneman III et al. 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 US examples. The study can thus serve to inform validation efforts in these contexts. In addition, it can inform the current discussion regarding the US 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 (US 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 sections present the research questions, methods and findings of the study. Finally, we draw conclusions and suggest further research.

Contextual background

The Chilean educational system is decentralised and consists of three types of schools: municipal (public), private subsidised and private non-subsidised. In 2008, there were approximately 11,907 schools working in the system, 49% of which were municipal schools, 44% private subsidised schools and 6% private non-subsidised schools (Ministry of Education 2009). Municipalities administer municipal schools, while private stakeholders (either individuals or private institutions) manage both private subsidised and private non-subsidised schools.

In 2008 Chile had roughly 176,500 classroom teachers, of which 55% worked in municipal schools (Ministry of Education. Departamento de Estudios y Desarrollo de la División de Planificación y Presupuesto del Ministerio de Educación de Chile 2010). 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 has been mandatory for teachers in municipal schools nationwide. For more details on the development of the teacher evaluation system and its characteristics see Avalos and Assael (2007) as well as Manzi et al. (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 (upwards and downwards) 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 summarised reports. The NTES 2008 results show that the majority (63.9%) of evaluated teachers received the performance categorisation 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 in earlier years. In total, until 2008, nine teachers had to leave the public teaching force due to unsatisfactory performance, but for some more their unsatisfactory evaluation result seems to have had a signalling effect, leading them to abandon the public education sector (Taut, Santelices, and Valencia 2010).

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, and 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' (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 categorisation of teacher performance. 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 convergent validity evidence 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 and Dempsey-Atwood 1976; Shavelson, Webb, and 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 US (e.g., INTASC, Praxis) include a combination of the following data sources: (a) some collection of teaching materials related to planning, instruction, and 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, and 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, and 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, and Burstein 1986; Darling-Hammond 1999);
(2) how to measure and analyse student learning gains so that they can validly and reliably reflect differences in teacher performance (e.g., Lissitz 2005; Kupermintz 2003; McCaffrey et al. 2003); and
(3) whether results of such value-added analyses should be included in teacher evaluation systems (Gordon, Kane, and Staiger 2006; Odden 2004; Braun 2005; Wright, Horn, and 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 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 programmes. For example, we closely studied the design of a construct and consequential validity study of the National Board for Professional Teaching Standards (NBPTS) accreditation process (Bond et al. 2000), and other NBPTS related validity studies (e.g., Goldhaber and Anthony 2004; Pool et al. 2001). Pecheone and Chung (2006) examined different aspects of the validity of The Performance

Assessment for California Teachers (PACT), while Le and Buddin (2005) provide a general overview of validity evidence for California teacher licensure exams. Wilson and Hallam (2006) 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 programme related to the validity of the Chilean national teacher evaluation system (NTES). This convergent validity study aimed at validating the final, high-stakes categorisation 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 and student learning gains 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 categorisations 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 the 2005 evaluation instruments and the 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 either an '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 tried to restrict their participation to the subject area for which we had standardised student achievement tests available (Mathematics). We also asked the teachers not to 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, audiotape a 1.5 hour lesson and take notes (classes were not video-taped);[1]
- Collect planning, instructional and student evaluation materials (including actual student work) pertaining to a two-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; and
- 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 the actual classroom performance of the participating teachers as possible. Teacher performance was conceptualised based on the same set of teaching standards (MBE) 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.

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 modelling in order to estimate the effect of teacher quality on students' standardised 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 them to take professional development courses, and to be re-evaluated 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 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 generalisability 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 incentives we offered.

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 rater training and 13% 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 in 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', 27 received NTES portfolio scores of 'competent' and 1 received a portfolio score of 'basic'. As is generally the case in NTES, our participants' final categorisation 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, 22 received portfolio scores of 'unsatisfactory', 1 received a portfolio score of 'basic' and 2 did not submit the portfolio. The local evaluation commission modified the final standing of the five remaining teachers included in the sample: four were raised from 'competent' (based on their performance on the instruments) to 'outstanding' and one was demoted from 'basic' to 'unsatisfactory'.

Students

A total of 1044 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 socio-economic categories. However, in the schools serving students from mid-high socio-economic backgrounds, we were able to recruit only 'outstanding' teachers.

Descriptive analyses and group comparisons

Teachers' standardised subject and pedagogical knowledge test

Overall the teachers' performance on the standardised subject and pedagogical knowledge test was poor. We think this may be in part explained by the fact that the test we used is part of a programme designed to certify teachers' pedagogical excellence. Although it is based on the same standards used by NTES, both programmes 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 a moderate practical significance of these statistical differences.

Classroom observations

The 25 items from the *post-observation questionnaire* 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 behaviour; 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 behaviour 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* show that 'outstanding' teachers spent slightly (but not significantly) more time on activities that are directly related to subject matter content and had 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 blindly 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. 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. 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 be more reflective on 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 classroom assessment 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).

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 out of the 22 group comparisons were 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.

Table 1. Statistical significance of group comparisons.

	t-test of equal means								
Indicator	Mean Outstanding Teachers	Std Dev	N	Mean Unsatisfactory Teachers	Std Dev	N	*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*

Table 1. (*Continued*).

	t-test of equal means								
Indicator	Mean Outstanding Teachers	Std Dev	N	Mean Unsatisfactory Teachers	Std Dev	N	*t*	Degrees of freedom	Sig. (bilateral)
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**

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

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

Students' standardised tests

The students' standardised tests were administered in 10 second-grade classrooms, 16 third-grade classrooms, and 14 fourth-grade classrooms. In both third and fourth grades we used a different form for the pre- and post-test, but this was not possible for second grade since there were too few items available. The pre-test was administered to a total of 1204 students and the post-test was administered to a total of 1160 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 = 1044) 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.[2] 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 were no missing data for students.

When comparing the 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 the post-test scores of these classrooms. Also, while we observe no statistically significant difference between the learning gains of 'outstanding' and 'unsatisfactory' teachers' students when aggregating them at the teacher (classroom) level, we do find statistically significant differences in the expected direction when comparing the learning gains of all students who were taught by 'outstanding' teachers with the learning gains of students who were taught by 'unsatisfactory' teachers.

The results from the hierarchical linear modelling analysis on the importance of teacher quality on student achievement as measured by our standardised tests are not conclusive as they vary depending on the model used.

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 this section. Models 1 and 3 were unconditional models.

Following Raudenbush and Bryk (2002), the models were estimated using group-mean centring because of the important variation observed in the mean of the level-1 covariates by teacher (level-2 unit).[3] 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' socio-economic characteristics.

Model 2

$$Achievement_{ij}^{2} = \beta_{0j} + \beta_{1j} * Achievement^{1}{}_{ij} + r_{ij}$$

$$\beta_{0j} = G_{00} + G_{1j} * TeacherQuality + u_{0j}$$

$$\beta_{1j} = G_{10} + u_{1j}$$

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 a 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 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. 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 modelled as a random variable with error u_{1j}. The three error terms are assumed to have a normal distribution with a mean of zero and variance σ_{rr}, σ_{u0}, σ_{u1} respectively.

In Model 4 the dependent variable was changed to represent 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

$$AchievementGrowth_{ij}^{2} = \beta_{0j} + \beta_{1j} * Achievement^{1}{}_{ij} + r_{ij}$$

$$\beta_{0j} = G_{00} + G_{1j} * TeacherQuality + u_{0j}$$

$$\beta_{1j} = G_{10} + u_{1j}$$

Teacher quality as measured by the NTES is statistically significant ($p = 0.01$) when achievement at post test is used as the outcome variable (model 2) but it is not statistically significant when using the difference in achievement (gain/growth/learning) as the level-1 outcome variable (model 4, p = 0.12).[4] 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, decrease = 14.34) and model 4 to the unconditional model (model 3, decrease = 8.02), 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 (Model 2 Deviance = 8734.65, Model 4 Deviance = 8654.41). However, we observe much smaller level-2 variance components for the gain score models (567.8 in Model 3 and 501.6 in Model 4) as compared to Models 1 and 2 (54.5 and 52.6 respectively), which indicates that between-teacher differences are much smaller in terms of students' gain scores than in terms of students' post-test scores.

The fixed effects estimated in Models 2 and 4 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 and 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 2). 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 standardised teacher knowledge test indicate that the difference between 'outstanding' and 'unsatisfactory' teachers is only of moderate importance.

Table 2. 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 behaviour	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

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. Only correlations larger than 0.3 are reported.

From the data collected through the classroom post-observation questionnaire, correlations are particularly strong between lesson structure as well as appropriate student behaviour and the NTES portfolio scores (r = 0.53 and 0.49 respectively, p = 0.00 in both cases). NTES' final category correlates moderately with lesson structure (r = 0.39, p = 0.00), appropriate student behaviour (r = 0.53, p = 0.00), and the proportion of time in which 95% of students are on task (r = 0.49, p = 0.00). Lower, but nevertheless statistically significant correlations were also observed with especially stimulating instructional practices of the teachers (r = 0.30).

The indicators of the teaching materials binder that most strongly correlate with the NTES instruments are those based on student evaluation design (r = 0.56, p = 0.00), student performance (r = 0.33, p = 0.00) and instructional materials (r = 0.52, p = 0.00). The correlations between the performance on NTES' instruments and students' and teachers' standardised 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 standardised tests do not correlate with any of the NTES instruments, but performance on the post-test correlates significantly with the peer evaluation (r = 0.41, p = 0.00) and the supervisor assessment (r = 0.41, p = 0.00). 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 (r = 0.41 and 0.31 respectively).

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 second grade and third grade standardised student tests, and the multiple-choice section of the standardised 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 hypothesise 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 due to the fact that the sections largely lacked materials (and that therefore the variance of scores in these sections is close to zero). The fourth grade math pre-test and the fourth 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 the second and third 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 behaviour). 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 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 behaviour, design of classroom assessment 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' communications 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 behaviour, students' time on task, and design of classroom assessment 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' standardised 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 classroom assessment design).

The comparisons that were not statistically significant have alternative explanations that are worth considering. For example, the pedagogical knowledge section of the teachers' standardised 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 standardised 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 analysed in this study. The student standardised 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 standardised 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 this is 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 teachers, 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, mid-career and experienced teachers. This framework has since been used in US school districts where standards-based teacher evaluation systems have been installed (see Odden and Kelley 2002). Danielson's framework includes 22 teaching standards organised into four domains: planning and preparation, classroom environment, instruction, and professional responsibilities. In Chile, a national consultation of teachers took place in which 80% of respondents validated the Chilean adaptation of these standards (Avalos and Assael 2007). Since then they have formed 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. 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 et al. 2010). 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 evidence of the NTES classification.

Our analyses show that both 'outstanding' and 'unsatisfactory' teachers perform well in planning their classes and designing instructional 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, seem to be weak in both 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 classroom assessment 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 teaching practice in the classroom and when rating their teaching materials. There were also statistically significant differences between 'outstanding' and 'unsatisfactory' teachers in their performance on the standardised subject knowledge test, but the small mean difference between the two groups seemed of less pedagogical importance than those differences observed in the binder and classroom observation, as confirmed by the effect sizes. We found inconclusive evidence of differential student learning (as measured by standardised tests) associated with 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 (Taut et al. 2010)? 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 allow 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.

Notes

1. 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.
2. 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.
3. 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.
4. The fact that teacher quality is statistically significant is especially important if we consider the size of the sample: 39 teachers and 1044 students.

Notes on contributors

Maria Verónica Santelices is an assistant professor at the Department of Education at Pontificia Universidad Católica de Chile. She received her PhD from the Graduate School of Education at the University of California, Berkeley.

Sandy Taut received her PhD from the Graduate School of Education at the University of California Los Angeles (UCLA). She currently leads the research unit at the Measurement and Evaluation Center (MIDE UC) of the School of Psychology at Pontificia Universidad Católica de Chile.

References

Avalos, B., and J. Assael. 2007. Moving from resistance to agreement: The case of the Chilean teacher performance evaluation. *International Journal of Educational Research* 45, nos. 4–5: 254–66.

Bond, L., T. Smith, W. Baker, and J. Hattie. 2000. *The certification system of the National Board for Professional Teaching Standards: A construct and consequential validity study.* Greensboro, NC: The University of North Carolina at Greensboro.

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.* Seattle, WA: University of Washington, Center for the Study of Teaching and Policy.

Goldhaber, D., and E. Anthony. 2004. *Can teacher quality be effectively assessed?* Washington, DC: The Urban Institute.

Gordon, R., T. Kane, and D. Staiger. 2006. *Identifying effective teachers using performance on the job.* Washington, DC: Brookings Institution.

Heneman III, H., A. Milanowski, S. Kimball, and A. Odden. 2006. *Standards-based teacher evaluation as a foundation for knowledge- and skill-based pay*. CPRE Policy Briefs RB-45. Philadelphia, PA: Consortium for Policy Research in Education, University of Pennsylvania, Graduate School of Education.

Hojat, M., and G. Xu. 2004. A visitor's guide to effect sizes. *Advances in Health Science Education* 9, no. 3: 241–9.

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, DC: 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, CA: Corwin.

Kane, M.T. 2001. Current concerns in validity theory. *Journal of Educational Measurement* 38, no. 4: 319–42.

Kupermintz, H. 2003. Teacher effects and teacher effectiveness: A validity investigation of the Tennessee value added assessment system. *Educational Evaluation and Policy Analysis* 25, no. 3: 299–318.

Le, V., and R. Buddin. 2005. *Examining the validity evidence for California teacher licensure exams.* Working Paper WR-334-EDU. Santa Monica, CA: RAND.

Lissitz, R., ed. 2005. *Value added models in education: Theory and applications.* Maple Grove, MN: JAM Press.

Manzi, J., D. Preiss, R. Gonzalez, P. Flotts, and Y. Sun. 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, in New York City, USA.

McCaffrey, D., J.R. Lockwood, D. Koretz, and L. Hamilton. 2003. *Evaluating value added models for teacher accountability.* Santa Monica, CA: RAND.

Messick, S. 1994. The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher* 23, no. 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. Philadelphia, PA: 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, CA: Corwin Press.

Ministry of Education. 2004. *Marco para la buena enseñanza* [Guidelines for good teaching]. Santiago: Ministerio de Educación.

Ministry of Education. 2009. Estadísticas de la Educación 2008. Departamento de Estudios y Desarrollo de la División de Planificación y Presupuesto del Ministerio de Educación de Chile. [Educational Statistics for 2008. Prepared by the Research and Development Area of the Budget and Planning Division of the Ministry of Education]. http://w3app.mineduc.cl/mineduc/ded/documentos/Estadisticas_2008_Capitulo_3.pdf (accessed December 24, 2010).

Ministry of Education. Departamento de Estudios y Desarrollo de la División de Planificación y Presupuesto del Ministerio de Educación de Chile. 2010. Sistema de Información de Estadísticas Educativas SIEE. [Information System about Educational Statistics]. http://w3app.mineduc.cl/Sire/index (accessed December 24, 2010).

National Research Council. 2008. *Assessing accomplished teaching: Advanced-level certification programs.* Washington, DC: National Academic Press.

Nye, B., S. Konstantopoulos, and L. Hedges. 2004. How large are teacher effects? *Educational Evaluation and Policy Analysis* 26, no. 3: 237–57.

Odden, A. 2004. Lessons learned about standards-based teacher evaluation systems. *Peabody Journal of Education* 79, no. 4: 126–37.

Odden, A.R., and C.J. Kelley. 2002. *Paying teachers for what they know and do.* 2nd ed. Thousand Oaks, CA: Corwin Press.

Pecheone, R.L., and R. Chung. 2006. Evidence in teacher education. The Performance Assessment for California Teachers (PACT). *Journal of Teacher Education* 57, no. 1: 1–15.

Peterson, K. 2000. *Teacher evaluation.* 2nd ed. Thousand Oaks, CA: Corwin Press.

Pool, J.E., C. Ellett, S. Schiavone, and C. Carey-Lewis. 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, no. 1: 31–48.

Porter, A.C., P. Youngs, and A. Odden. 2001. Advances in teacher assessments and their uses. In *Handbook of research on teaching*, ed. V. Richardson, 259–97. Washington, DC: American Educational Research Association.

Raudenbush, S., and A. Bryk. 2002. *Hierarchical linear models. Applications and data analysis methods.* London/New Delhi/Thousand Oaks, CA: Sage.

Shavelson, R., and N. Dempsey-Atwood. 1976. Generalizability of measures of teaching behavior. *Review of Educational Research* 46, no. 4: 553–611.

Shavelson, R., N. Webb, and L. Burstein. 1986. Measurement of teaching. In *Handbook of research on teaching,* 3rd ed., ed. M. Wittrock, 569–98. New York: Macmillan.

Shulman, L.S. 1987. Knowledge and teaching: Foundations of the new reform. *Harvard Educational Review* 57, no. 1: 1–22.

Taut, S., V. Santelices, C. Araya, and J. Manzi. 2010. The theory underlying a national teacher evaluation program. *Evaluation and Program Planning* 33, no. 4: 477–86.

Taut, S., V. Santelices, and E. Valencia. 2010. *Resultado de reevaluaciones y situación laboral de los docentes evaluados por el Sistema de Evaluación de Desempeño Docente entre 2003 y 2008* [Subsequent performance and employment situation of teachers in the National Teacher Evaluation System]. Informe Técnico [Technical Report] IT 1007. Santiago: MIDEUC. http://www.mideuc.cl/docs/informes/it1007.pdf (accessed December 24, 2010).

US Department of Education. 2009. Overview information; Race to the Top Fund; Notice inviting applications for new awards for fiscal year 2010. http://www2.ed.gov/programs/racetothetop/applicant.html (accessed January 31, 2010).

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., and P. Hallam. 2006. 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, April, in San Francisco, CA.

Wright, P., S.P. Horn, and W. Sanders. 1997. Teacher and classroom context effects in student achievement: Implications for teacher evaluation. *Journal of Personnel Evaluation in Education* 11, no. 1: 57–67.