education policy analysis archives

A peer-reviewed, independent,
open access, multilingual journal

Arizona State University

Volume 22 Number 71 July 21st, 2014 ISSN 1068-2341

The Development and Implementation of a National, Standards-based, Multi-method Teacher Performance Assessment System in Chile

Sandy Taut
&
Yulan Sun
Pontificia Universidad Católica de Chile
School of Psychology
Measurement Center MIDE UC
Chile

Citation: Taut, S. and Sun, Y. (2014). The Development and Implementation of a National, Standards-based, Multi-method Teacher Performance Assessment System in Chile. *Education Policy Analysis Archives, 22* (71). http://dx.doi.org/10.14507/epaa.v22n71.2014.

Abstract: Assessing teaching performance is a contentious issue in current educational policy in many countries. In Chile, a national, standards-based, multi-method, mandatory teacher evaluation system has been in place since 2003 to assess the performance of about 70, 000 public school teachers from pre-school to high school and adult education. The Chilean system combines formative and summative purposes and uses four instruments: a structured portfolio, a peer interview, supervisor questionnaires, and a self-assessment. In this paper we analyze the Chilean system as a case of interest regarding some controversial issues in teacher evaluation today. To this end we describe the system's political background, implementation process and results. We put a special focus on the development and evolution of its evaluation instruments; we discuss their strengths and limitations and review the research evidence regarding their validity and reliability. Finally, we reflect on the Chilean experience in terms of the insights it can provide into teacher performance assessment for an international audience.

Journal website: http://epaa.asu.edu/ojs/
Facebook: /EPAAA
Twitter: @epaa_aape

Manuscript received: 8/19/2013
Revisions received: 12/17/2013
Accepted: 1/11/2014

Keywords: Teacher performance assessment; teacher evaluation; standards; evaluation instruments; portfolio; validity; measurement system implementation; Chile.

Desarrollo e Implementación de un Sistema Nacional de Evaluación de Profesores, multi-método y basado en estándares, en Chile.

Resumen: La evaluación del desempeño docente es hoy un tema candente dentro de las políticas educacionales en muchos países. En Chile, desde 2003 se ha aplicado un sistema nacional y obligatorio de evaluación de profesores, basado en estándares y que utiliza múltiples instrumentos y fuentes. El programa ha evaluado aproximadamente a 70, 000 profesores de escuelas municipales (públicas) que ejercen en distintos niveles educacionales, desde pre-escolar a educación secundaria y de adultos. El sistema chileno combina propósitos formativos y sumativos y utiliza cuatro instrumentos: un portafolio estructurado, una entrevista por un par, cuestionarios a directivos y una autoevaluación. En este artículo analizamos el sistema chileno como un caso de interés en relación con algunos puntos altamente controversiales en evaluación docente hoy. Para ello, describimos el contexto político del sistema, el proceso de implementación y sus instrumentos y resultados, poniendo especial énfasis en el desarrollo y evolución de los instrumentos de evaluación; discutimos las fortalezas y limitaciones de éstos y revisamos la evidencia de investigación disponible en torno a su validez y confiabilidad. Finalmente, reflexionamos en torno a la experiencia chilena intentando extraer de ella algunas luces acerca de la evaluación del desempeño docente para una audiencia internacional.

Palabras-clave: Evaluación de profesores; evaluación docente; instrumentos de evaluación; portafolio; validez; implementación de sistemas de medición; Chile.

Desenvolvimento e implementação de um Sistema Nacional de Avaliação de Professores com múltiplas metodologias e baseados em padrões no Chile.

Resumo: A avaliação do desempenho docente é hoje um tema candente nas políticas educacionais em diversos países. Em Chile, desde 2003, aplica-se um sistema nacional e obrigatório de avaliação de docentes, baseado em padrões e que utiliza múltiplos instrumentos e fontes de dados. O programa avaliou aproximadamente 70 mil professores de escolas públicas de distintos níveis de educação, desde da Educação Infantil (pré-escola) até o Ensino Médio e Educação de Jovens e Adultos (EJA). O sistema chileno de avaliação combina propósitos formativos e somativos, utilizando-se de quatro instrumentos, a saber: um portofólio estruturado, uma entrevista realizada por um par, questionários para os diretores das escolas e uma autoavaliação. Neste artigo analisamos o sistema chileno como um estudo de caso para alguns pontos que são altamente controversos na discussão atual de avaliação docente. Para isso descrevemos o contexto político do sistema de avaliação, o seu processo de implementação, seus instrumentos e resultados, enfatizando o desenvolvimento e evolução dos instrumentos de avaliação. Discutimos, também, as fortalezas e limitações dos instrumentos e revisamos a evidencia de investigação disponível em relação da validez e confiabilidade dos mesmos. Finalmente, refletimos em torno da experiência chilena no sentido de trazer alguma luz sobre a avaliação do desempenho docente para um público internacional.

Palavras-chave: Avaliação de professores; avaliação docente; instrumentos de avaliação; portifólio; validez; implementação de sistemas de medição; Chile.

Introduction[1]

Teacher evaluation is a controversial and central issue in current educational policy in many countries. This is evidenced, for example, by a recent series of reviews of national teacher evaluation systems commissioned by the Organization of Economic Cooperation and Development (OECD, 2013a, 2013b). Aside from its political complexities, designing and implementing teacher evaluation systems involves conceptual and methodological issues that have no single answer within literature and empirical research. Three of them seem particularly important. First, in terms of what should be assessed, defining teaching quality is conceptually and methodologically complex because of the nature of teaching as an interaction among teacher, students, content, and context (Cohen, Raudenbush, & Ball, 2003; Rose, 2011) and the fact that conclusive empirical evidence on effective teaching practices is still in its infancy (Hattie, 2009). There is no single, accepted conceptual model about what is competent teaching and the solutions also reflect value-based discussions and differ across contexts (Kennedy, 2010).

Second, the most appropriate methods to evaluate teaching quality have been under strong discussion. Methodological options include teacher tests (e.g., on subject matter knowledge), teacher surveys (e.g., on the frequency of certain classroom practices, or beliefs and dispositions), as well as student surveys, classroom observation (either direct or using videos), portfolios that ask the teacher to collect teaching materials, evidence of student learning, and to answer reflective questions, and value-added methods that determine the "teacher effect" based on student learning on standardized achievement tests (Martínez, 2012). At a conceptual level there is no consensus about which method is most appropriate and what weight each one should have if combined. Depending on the political and social context in which teaching is immersed, educational jurisdictions have opted for different methods of providing teachers with feedback on their performance and to hold them accountable (Isoré, 2009).

Finally, another important aspect is the purpose that teacher evaluation is designed to meet. This involves an accountability (summative) function and an improvement or developmental (formative) purpose, and usually some combination of both. In the first case, the key concern is to distinguish between effective and ineffective teachers, with positive (e.g. economic incentives) and negative (e.g. dismissal) consequences attached to the results. In a formative approach the main goal is to help teachers improve their performance through in-depth diagnosis and feedback. Currently discussions reflect a tension between an accountability-oriented approach in which the judgment regarding teacher effectiveness includes students' standardized test results, among other sources of information, and an approach focusing mainly on teachers' professional development, usually excluding student test scores and using classroom observations as a main source of information.

Responses to these three dimensions have been varied depending on the respective political and historical context, and internationally we find considerable heterogeneity of programs, methods and instruments (Isoré, 2009; OECD, 2013a, 2013b). In this paper, we intend to contribute to the international discussion through the analysis of the National Teacher Evaluation System (NTES, also known as "Docentemás") implemented in Chile since 2003. Several reasons make this an especially interesting case. First, unlike other educational systems, particularly the United States, where Race to the Top legislation now requires states to include student learning as an indicator in teacher evaluation, in Chile student learning is not included as a direct indicator in teacher evaluation. Instead, the Chilean system operationalizes a set of teaching standards that correspond to

[1] [1] The authors acknowledge FONDECYT No. 1120441 for partial financial support.

[2] For a complete example of this report (in Spanish), go to http://www.docentemas.cl/pageflip/001b/index.html

[3] Approximate cost is based on 2012 data: the total amount allocated in the national budget for the annual

descriptions of what competent teaching should look like (based on Danielson, 1996). Also, the Chilean NTES represents an example of the mix of methods and evaluator perspectives that recent research on teacher evaluation seems to call for (Darling-Hammond, 2012; Kennedy, 2010; OECD, 2013a, 2013b). It includes a combination of instruments and the perspectives of peers, supervisors, and teachers themselves. Furthermore, the NTES combines formative and summative purposes.

The NTES has been implemented for almost 10 years, resulting in a wealth of information that is interesting to share at the international level. This includes not only the experience and documentation regarding the installation and implementation process, but also a comprehensive research agenda regarding the validity and reliability of the instruments and the program's uses and consequences (Alvarado, Cabezas, Falck, & Ortega, 2012; León, Manzi, & Paredes, 2008; Orellana, & Merino, 2013; Taut, Santelices, & Stecher, 2012). Finally, a recent in-depth study commissioned by the Ministry of Education to the OECD has just been released and offers an external, international expert review of the system (Santiago, Benavides, Danielson, Goe, & Nusche, 2013).

We briefly describe our objectives and methods, present an overview of the Chilean educational context, characterize the Chilean NTES, and describe the evaluation instruments, also analyzing their strengths and weaknesses. Finally, based on our experience and existing research we discuss some lessons that can inform the improvement of this and other teacher evaluation systems, thus contributing to the current debates around teacher evaluation.

This paper represents two perspectives: both authors belong to the university that has acted as a consultant to the Ministry of Education for the implementation of the NTES, but one in charge of the team of professionals responsible for instrument design and assessment implementation, and the other as an independent academic researcher who developed a considerable part of the research agenda associated to the NTES. The joint venture between university researchers and measurement professionals in charge of the program has the advantage of direct access to internal documentation and first-hand knowledge of processes and outcomes on the one hand, while on the other hand safe-guarding standards of sound scientific production and ethical conduct (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 1999).

Objectives and Methods

Objectives

The purpose of the current paper is to describe and analyze the Chilean national teacher evaluation system (NTES) with a particular focus on its evaluation instruments. We discuss the strengths and weaknesses of each instrument, as well as the system as a whole, based on the empirical evidence regarding its quality and the authors' in-depth knowledge of the system.

Methods

In order to meet the objectives detailed above we reviewed existing evidence and publications regarding the Chilean national teacher evaluation system. We searched a total of 21 local and international databases, including papers since 2008 and using the following search terms (in English and Spanish): evaluation of teaching, teacher assessment, teacher evaluation, Chilean, Chile. As a result 22 articles alluding to the Chilean NTES were found; 17 of them were in fact directly related to the NTES. These included conceptual reviews, quantitative empirical studies that used NTES data, qualitative studies regarding the uses and consequences of the NTES, and validation research studies. Of these 17 papers, ten articles correspond to empirical studies carried out by researchers of the Catholic University of Chile. In fact, most references regarding the

program's validation are co-authored by one of the authors of this paper. It is important to mention that a large part of this research was produced with external funding based on peer-review processes; papers were published in peer-reviewed, international journals. In addition, it is well-documented by existing standards of educational measurement and evaluation that assessment developers themselves have a main responsibility of supporting validation research (AERA, APA, & NCME, 1999; Centro Nacional para la Educación Superior [CENEVAL], 2000; Educational Testing Service [ETS], 2002). In fact, we believe that it constitutes one of the main strengths of the Chilean national teacher evaluation system that it counts with a considerable body of validation evidence.

The Chilean Educational Accountability Context

In educational circles around the world, Chile is known for its free-market approach to education. In addition, over the last two decades, educational accountability has gained momentum in Chile. Accountability policies have focused on monitoring student achievement and evaluating teacher performance. Chile has one of Latin America's oldest and most sophisticated student testing systems (Sistema de Medición de la Calidad de la Educación, SIMCE) and uses test results to hold schools accountable (e.g., Sistema Nacional de Evaluación de Desempeño, SNED). A new national-level institution (Agencia de Calidad) is in charge of monitoring school quality starting in 2013 and will be using test results and other educational indicators to classify schools into one of four levels of quality. For many years, test results have been published in the form of school rankings by the country's leading newspapers.

Since the early 2000s Chile has been implementing teacher accountability policies based on technically complex evaluation measures such as portfolios (Manzi, González, & Sun, 2011b). In 2002 Chile started implementing a teaching excellence certification and incentive system (Asignación de Excelencia Pedagógica, AEP) and, one year later, the NTES. The political process of introducing the NTES started in the 1990s and was characterized by difficult negotiations with the most important political stakeholders and resistance from a considerable segment of teachers. During the military regime (1973-1989), because teachers were seen as an important opposition force, teacher education was removed from universities and relegated to a technical career with low pay and benefits. In the 1990s, the new democratic government tried to address concerns about teachers and teaching quality. The resulting initiatives first intended to reinstall teaching as a more attractive career, for example, by increasing salaries and making teachers public servants with lifetime tenure, and later introduced the logic of performance-based accountability. Despite concerns and resistance on the part of teachers, in 2002 a committee consisting of teacher union representatives, representatives of local municipal authorities, and Ministry of Education personnel arrived at a consensus to conduct performance-based teacher evaluation in roughly its current form (Assael, & Pavez, 2008; Avalos, & Assael, 2006).

Teacher Evaluation in Chile: An Overview

General assessment system description

The evaluation is mandatory for classroom teachers of public (municipal) schools and distinguishes four levels of performance: "outstanding, " "competent, " "basic, " and "unsatisfactory". Teachers must complete the evaluation every four years if in the first two categories, every two years if their result has been "basic, " and the following year if it was "unsatisfactory." Evaluation instruments include 1) a structured portfolio comprising a written part and a videotaped lesson, 2) a peer interview, 3) a supervisor assessment, and 4) a self-assessment (for

sample instruments in Spanish, see http://www.docentemas.cl). The scores for each instrument have different weights in the final performance categorization, as defined by law: the portfolio assessment contributes 60% of the final score, peer interview 20%, and supervisor and self-assessment 10% each. These are combined for an overall result, which is then ratified or modified by a local evaluation commission constituted by the municipal peer evaluators and the municipal educational authority (the latter only with voice but no right to vote).

The NTES is standard-based, following the guidelines established in the *Marco para la Buena Enseñanza*, or Framework for Good Teaching (Ministry of Education, 2004). The instruments cover different aspects of competent teaching as defined by this framework. The evaluated teachers receive a descriptive report detailing their results for the portfolio dimensions (see Table 1) as well as the other three instruments[2]. Essentially, the report describes the results of the portfolio assessment, including the teacher performance for each dimension and each indicator (see Table 2 for an example), the aggregated results of the peer interview and supervisor assessment, the results of the self-evaluation, and the overall performance level.

Table 1.

Dimensions and indicators of the 2011 portfolio

	Dimensions	Indicators
Module 1 (implementing a learning unit)	A. Unit organization	- Formulation of learning objectives - Relationship between learning objectives and activities - Unit sequence
	B. Analysis of unit lessons	- Analysis of students' characteristics - Analysis of unit's highly effective elements - Analysis of unit's less effective elements
	C. Quality of classroom assessment	- Classroom assessment instrument - Grading or scoring guidelines - Relationship between learning objectives and classroom assessment
	D. Analysis of students' classroom assessment results	- Responsibility taken for student learning - Feedback given to a student
	E. Pedagogical reflection	- Reflection about students' learning difficulties - Reflection about students' motivation - Analysis of in-service training needs as a teacher
Module 2 (videotaped lesson)	F. Classroom learning environment	- Classroom environment - Promoting students' participation in the lesson - Accompanying students' activities
	G. Lesson structure	- Quality of the lesson opening - Quality of the lesson closure - Contribution of classroom activities to learning objectives
	H. Pedagogical interaction	- Quality of explanations - Quality of questions - Quality of feedback given to students - Implementing teaching strategies specific to grade level and subject curriculum

[2] For a complete example of this report (in Spanish), go to http://www.docentemas.cl/pageflip/001b/index.html

Note: While the dimensions are relatively stable over time, some indicators change from year to year. Also, since 2012, the indicators evaluated in Module 1 were re-organized into 4 dimensions instead of 5.

Table 2
Example of feedback provided for one dimension in 2012 Individual Report

Dimension C: Quality of classroom assessment.		
Indicators	Result	Performance description
Classroom assessment instrument and scoring	Competent	The instructions, questions, or tasks included in the assessment are clear and correct. Also, in your scoring guidelines, the expected performance or responses are correctly described or identified.
Relationship between learning objectives and instrument	Basic	The assessment instrument only partially addresses the learning objectives, since not all of them are evaluated, or some of the assessment questions, items, or situations are not related to any objective.

Note: A similar table is presented for all the indicators grouped by each dimension.
Source: http://www.docentemas.cl/pageflip/001b/index.html

The school principal and the municipal education authority also receive reports providing general information (final result and portfolio result) about all the classroom teachers in their school or municipality, and more detailed data for the specific group of evaluated teachers in that year (group results for each portfolio dimension with a short description of each dimension).

Legal framework and consequences

The Chilean national teacher evaluation system (NTES) was introduced gradually by the Ministry of Education in 2003–2004 in a few municipalities, and since 2005 has been fully implemented at the national level and is mandatory for teachers in municipal schools.

An important factor in the introduction of the NTES was the previous existence of a set of professional teaching standards, the Framework for Good Teaching (FGT) describing what good teaching performance should look like in the context of initial teacher licensure (see Fig. 1). The FGT is based on Danielson's Framework for Teaching (Danielson, 1996), which is also used for teacher evaluation in various school districts in the United States (Heneman, Milanowski, Kimball, & Odden, 2006) as well as in the Measures of Effective Teaching (MET) Project (Bill and Melinda Gates Foundation, 2011, 2012, 2013).

Teaching Preparation

A.1. Master the subjects taught and the national curricular framework.
A.2. Know the characteristics, knowledge and experiences of his/her students.
A.3. Master the didactics of the subjects or disciplines taught by him/her.
A.4. Organize the objectives and contents consistently with the curricular framework and the characteristics of particular students.
A.5. Use evaluation strategies that are consistent with the learning objectives, the subject taught, and the national curricular framework, and allow all students to show what they have learnt.

Creating a Learning Environment

B.1. Create an environment dominated by values such as acceptance, equality, trust, solidarity and respect.
B.2. Create an environment dominated by values such as acceptance, equality, trust, solidarity and respect.
B.3. Show high expectations about the learning possibilities and development of all of his/her students
B.4. Create and keep consistent regulations about classroom coexistence.
B.5. Create an organized working atmosphere and make available the spaces and resources required by the learning process.

A B
D C

Professional Responsibilities

D.1. Reflect systematically about his/her teaching skills.
D.2. Build a professional and team relationship with his/her peers.
D.3. Take up responsibilities regarding student counseling.
D.4. Promote respect and carry out cooperation actions with his/her students' parents and guardians.
D.5. Manage updated information relevant to the teacher profession, the educational system and the current policies.

Teaching for the Learning of Every Student

C.1. Be able to communicate the learning objectives in a clear and accurate way
C.2. Design challenging and consistent teaching strategies that are relevant for the students.
C.3. Treat the classroom contents with the right conceptual focus and using terms that students are able to understand.
C.4. Optimize the time available for teaching.
C.5. Promote the development of thought.
C.6. Evaluate and monitor the process of understanding and the appropriation of contents by the students

Figure 1. Overview of the Framework for Good Teaching (FGT).

The legal backing for the evaluation came in 2005, making evaluation legally binding and attaching negative consequences to the refusal to participate (Decree No. 192, 2005; Law 19.961, 2004). As mentioned before, the system combines summative and formative purposes. Teachers with basic and unsatisfactory performance have to undergo mandatory professional development (Planes de Superación Profesional or Professional Development Plans). Also, if unsatisfactory teachers repeat their below-acceptable performance in consecutive evaluations, dismissal is mandatory. Teachers with competent or outstanding performance, on the other hand, can take a

subject and pedagogical knowledge test and, if successful, receive an annual salary bonus. The size of the bonus depends on the combined performance from the evaluation and the test.

In 2011, Congress passed legislation that changed some of the consequences of the system (Law 20.501, 2011). For example, the number of consecutive unsatisfactory results leading to dismissal was reduced from three to two; teachers whose performance is assessed as "basic" are now also subject to possible removal from the classroom if they fail to elevate their performance to the "competent" level. In addition, principals can now remove 5% of their teachers based on "basic" or "unsatisfactory" NTES performance. Also, the bonus for high-performing teachers was increased.

The formative purpose of the system is explicitly mentioned in the corresponding law, referring to "a teacher evaluation system of a formative nature that is focused on improving the pedagogical work of educators and on promoting their continuous professional development" (Decree No. 192, 2004, p. 1). Therefore, the evaluation is intended not only to pass judgment at the individual teacher level but also to be formative by providing useful information at different levels of the educational system. For example, the results are supposed to inform educational decision-making and personnel decisions at the municipal level, and evaluated teachers are expected to use the evaluation process and results to reflect on their practice, work on overcoming diagnosed weaknesses (through targeted professional development), and maintain good practices. The evaluation is also meant to contribute to teacher peer collaboration by fostering conversations about good practice. Furthermore, the access to incentives for high-performing teachers is supposed to improve teachers' job commitment and satisfaction. The underlying stakeholder theory of the teacher evaluation system is described in detail in Taut, Santelices, Araya and Manzi (2010).

Implementation

Implementation of the evaluation system extends over a full year, with different stages starting with creating the list of teachers to be evaluated to communicating the final results. The following diagram illustrates the process and timeline (see Fig. 2).

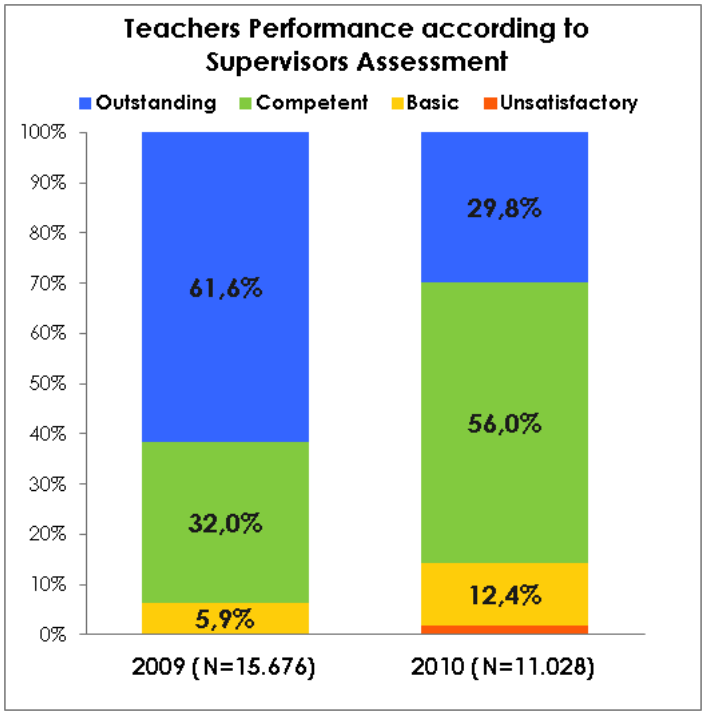

Figure 2. Supervisor questionnaire results 2009 versus 2010, after change was introduced.

An important aspect of implementation is that the evaluated teachers have easy access to information about the evaluation process. First, a website provides all the evaluation materials, frequently asked questions, background information, legal documents, and contact information regarding the evaluation. Second, a call center is available while the teachers are working on their portfolios to clarify logistical questions. Third, municipal authorities responsible for implementing the evaluation at the local level receive training and serve as a direct contact point for the teachers being evaluated.

Another important aspect of implementation is quality data entry and data security. Each year, the information contained in the evaluation instruments is entered in a software program especially designed for this purpose. The software then analyzes the data and generates reports for each teacher, school, and municipality. All professionals working on the evaluation sign confidentiality agreements and data used for research purposes are distributed only with permission from the evaluated teachers, signed as part of the background questionnaire evaluated teachers are asked to fill out as part of the portfolio.

In 2012, the cost for each evaluated teacher was about US$400 (CLP$200, 000).[3] Most of the expenses go to portfolio scoring because of the large number of specialized teachers hired for this task, video recording of the lessons, as well as the technical team in charge of developing the evaluation.

Results

The distribution of results shows a similar profile along the years. The majority of teachers obtained a "competent" result (52.4% to 64%, depending on the year), while almost a third received a "basic" performance assessment. Only a relatively small percentage is evaluated as "outstanding, " and very few are considered "unsatisfactory" (Sun, Correa, Zapata, & Carrasco, 2011; see Table 3).

Table 3
NTES results 2003–2010

Year	Unsatisfactory	Basic	Competent	Outstanding	Total N Evaluated
2003	3.7%	30.3%	56.6%	9.4%	3, 673
2004	3.0%	34.3%	52.4%	10.3%	1, 719
2005	3.5%	37.2%	52.7%	6.6%	10, 665
2006	2.8%	31.4%	59.0%	6.8%	14,190
2007	2.0%	33.0%	56.6%	8.5%	10, 413
2008	1.1%	22.8%	64.0%	12.1%	16, 015
2009	1.5%	28.9%	63.1%	6.5%	15, 699
2010	2.6%	33.3%	58.1%	6.0%	11, 061

Source: Manzi, González & Sun (2011), p. 96.

[3] Approximate cost is based on 2012 data: the total amount allocated in the national budget for the annual implementation of the NTES, divided by the number of evaluated teachers.

If we compare the results across evaluation instruments, the portfolio scores pull the distribution "down" toward lower performance, while the other instruments (especially the self-assessment) pull the distribution "up" toward higher performance (Ministry of Education, 2013; see Table 4).

Table 4
Mean scores of NTES 2009 –2012 by instrument

Year	Self-assessment	Supervisor reports	Peer interview	Portfolio
2009	3.88	3.15	2.93	2.21
2010	3.88	2.86	2.97	2.18
2011	3.86	2.93	3.15	2.20
2012	3.87	2.91	3.19	2.25
Average	3.87	2.96	3.06	2.21

Source: Ministry of Education (2013).

Evaluation instruments

In this central section of the paper, we describe the four evaluation instruments, which were developed by measurement experts at a university measurement center following the guidelines of the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999) and the *Standards for Personnel Evaluation* (Joint Committee on Standards for Educational Evaluation, 2009).

Portfolio development

The portfolio consists of two parts, or modules. For the first module, the teacher chooses one of two pre-determined learning objectives from the national curriculum, by subject and grade level, to design a teaching unit comprising eight lessons. The portfolio is standardized to make the evaluation experience comparable across subjects and grade levels. The teacher has to hand in materials from the teaching unit he or she actually implemented. These materials include lesson plans, materials reflecting pedagogical activities and classroom assessments, as well as written responses to questions related to how the teacher uses the materials, the student assessment results, and the capacity to reflect on his or her teaching practice. This corresponds to five dimensions of teaching practice, each further operationalized by three indicators (see Table 1).[4]

For the second module, one of the teacher's lessons (40 minutes) is videotaped by an external specialized contractor. The videotape gives insight into the teacher's interactions with the students, the classroom climate, and the actual implementation, structure, and activities of one particular lesson. The teacher knows ahead of time when he or she will be videotaped. Afterwards, the teacher completes a short questionnaire regarding the lesson (e.g., number of students, learning objectives, special circumstances). This second module covers three dimensions of teaching practice (see Table 1).

[4] Since 2012 the indicators have been reorganized into only four dimensions, but without changing the focus of the portfolio instrument.

The portfolios are developed according to the teaching standards (FGT) and additional sources of information:

- videos of teaching practice for each subject and grade level,
- detailed revisions of portfolio instructions by a team of expert teachers pertaining to the different subjects and grade levels, as well as technical experts,
- pilot studies of the draft portfolios, applying the think-aloud technique individually and in groups of teachers,
- pilot studies of the final version of the portfolios resulting in about N=12 completed portfolios for each subject and grade level, to be used for constructing the scoring rubrics and for rater training.

Teachers are given 12 weeks to complete the portfolio. A consistent and frequent concern voiced by teachers regarding the evaluation is the additional workload necessary for completing the portfolio, since they do not receive extra time as part of their contract to work on the evaluation. This is shown in the questionnaire that accompanies the portfolio, in which consistently over the years more than 70% of respondents reported lack of time as the principal difficulty in developing the portfolio. At the same time, however, teachers also consistently consider portfolio development a useful and relevant professional development activity, which allows them to reflect on and revise their teaching practice and interact in meaningful ways with their peers (Taut, Santelices, Araya, & Manzi, 2011).

Portfolio scoring

The portfolios are scored by trained raters using a detailed scoring rubric. While most indicators are general for all subjects and grade levels, some indicators are unique for each subject and grade level. As specified by law, the portfolio raters have to be in-service classroom teachers with at least five years of teaching experience. They only score portfolios for their subject area and grade level. The rubrics define each of the four performance categories for each indicator and are constructed following a process similar to the portfolio. The rubrics' dimensions and some of the indicators are maintained to ensure comparability over time, while other indicators are replaced in order to continuously improve the instrument, as well as to avoid cheating. Scoring rubric elaboration starts by defining "competent" performance, followed by defining "basic, " then "unsatisfactory, " and last "outstanding" performance. Operationalized definitions of "competent" performance are extrapolations of the teaching standards (FGT) and based on the professional judgment of the pedagogical and subject matter experts. The piloting of the rubrics provides information regarding the adequacy of the operationalization of the performance levels for each indicator. Gradually, rubrics have evolved from more specific and quantitative to more holistic, relying more on qualitative rater judgments. Table 5 shows an illustrative example of how the scoring rubric might operationalize an indicator.

Table 5
Illustrative example of operationalized indicator in portfolio scoring rubric

Dimension F: Classroom learning environment	**Indicator "Promoting students' participation in the lesson"**
Outstanding	Competent elements PLUS: The teacher encourages student participation in the lesson and explicitly recognizes the value of different opinions and answers as a source of enrichment for the learning process. OR If some students do not spontaneously participate in classroom activities, the teacher uses strategies or actions to promote student involvement.
Competent	The teacher offers opportunities for participation by all students, not just a few. AND All opportunities offered by the teacher for student participation are related to the learning objectives and/or content of the lesson.
Basic	The teacher offers opportunities for participation by all students, not just a few. AND Most of the opportunities offered by the teacher for student participation are related to the learning objectives and/or content of the lesson.
Unsatisfactory	Does not accomplish some of the Basic elements. OR Most of the time, the teacher does not answer students' questions during the lesson.

Note: The actual portfolio scoring rubrics used in the Chilean evaluation system are confidential. The example shows a preliminary version of a rubric used in the past.
Source: Manzi, González & Sun (2011), pp. 50–51.

All indicators have the same weight in the final score for each dimension, and all dimensions have the same weight in the final score of the portfolio. This decision was based on the fact that the underlying standards place equal emphasis on each aspect of good teaching. The measurement experts in charge of the NTES have questioned the practice of equal weighting and plan to establish with more authority whether to weigh specific aspects of teaching practice differently in calculating the final evaluation score (e.g., the evaluation of subject-specific pedagogy should perhaps receive more weight in the future).

In a typical year, the scoring process involves five rating centers (run by local universities) employing about 60 supervisors who supervise about 450 raters, in order to be able to complete scoring of about 15, 000 portfolios over a four-week period during summer vacations. It typically takes about 50 minutes to score a video, and between 50 and 60 minutes to score a written module.

In terms of assuring the quality of the rating process, all supervisors are trained for 64 hours, and all raters are trained for 24 hours. During the training they get to know the scoring rubric and learn to apply it on practice portfolios. After the initial training there is a trial period of three days of scoring when all processes are tested and raters are monitored while continuing to practice. This helps to identify supervisors and raters who diverge from the pre-established scores and makes it possible to replace them if low performance persists. In addition, at least a third of the raters every year have already worked as raters in previous years and therefore have obtained considerable expertise over time. Every Monday during the actual four-week scoring period, all raters participate in a group scoring session with their supervisors for purposes of re-calibration. In addition, 20% of randomly selected portfolios for each subject and grade level are double-rated. If the two raters differ substantially, then the supervisor functions as a third rater who resolves the discrepancies.

Strengths and weaknesses of portfolio development and scoring

Although the OECD review of the NTES (Santiago et al., 2013) suggests some changes to the portfolio so that it could represent more of a "natural harvest" of teachers' daily practice, given that the NTES is a mandatory, high-stakes national evaluation program we consider its level of standardization as a necessary characteristic. Along the same lines, in a 2013 report the OECD (2013b) recognizes that a certain level of standardization is unavoidable when a teacher appraisal system serves accountability purposes, in order to ensure more reliable and unbiased decision-making.

Another strength of the portfolio is the face validity it enjoys among teachers and the positive effects teachers report its elaboration process to have in terms of their professional development. In the background questionnaire teachers consistently report that the portfolio is a useful instrument that helps them reflect on and revise their practice, and that the portfolio evidence reflects their actual classroom practice (Sun et al., 2011b; also see Darling-Hammond, Wei, & Johnson, 2009; Hakel, Koenig, & Elliott, 2008).

Also, different studies have validated the portfolio based on its relation with students' test results. Alvarado et al. (2012) indicate that among the four NTES instruments, the portfolio shows the strongest relationship with student achievement as measured by standardized test results, followed by the supervisor assessment. Taut, Valencia, Santelices, Palacios, Jiménez and Manzi (2013) have used value added methodology to relate teachers' NTES scores with their students' learning progress over a two-year time span. They found that relations were strongest for teachers' portfolio scores, especially in mathematics, reaching correlation coefficients of about 0.3, depending on the model used.

On the other hand, one weakness of the portfolio is the possibility of copy or fraud, when teachers present materials that do not correspond to their actual practice[5]. Although it is unclear how much of a problem this is, the consequences of possible fraud or copy should be more explicit, more severe, and better enforced.

Internal consistency of the portfolio (as well as the other instruments) has been at acceptable levels with Cronbach Alpha around or above 0.8 (Taut, Santelices & Manzi, 2011). Factor analyses, on the other hand, have pointed to possible improvements. The way indicators are grouped into dimensions, which in turn are used to report results, could be changed to more accurately reflect the empirically determined underlying constructs. Results have varied somewhat over the years, but in

[5] Although there is no automated system for detecting cheating (because portfolios are delivered printed), each year a small number of cases of cheating is identified when two raters notice an excessive resemblance of two portfolios. Furthermore, a simple internet search shows offers regarding the elaboration of the portfolio in exchange for money.

general, the exploratory factor analysis identified either five or six factors for the entire portfolio, including the written part and the videotaped lesson. The factors associated with the videotaped lesson change from year to year, but in some years have neatly recreated the underlying theoretical dimensions. The factors associated with the written portfolio have been more stable over time. We call them (a) lesson planning, (b) designing classroom assessment materials, and (c) reflecting on pedagogical decisions. These empirical factors combine similar tasks that are repeated across dimensions, for example, asking teachers to reflect on the work included in the portfolio (reflecting on lesson planning as well as classroom assessments, for example). The results from the confirmatory factor analysis (conducted using 2010 portfolio data) indicate that the portfolio's theoretical structure of the eight dimensions fits the data well (for more details, see Taut et al., 2012).

Since 2005 generalizability studies have been conducted to examine the percentage of variability in the portfolio scores that was attributable to (a) a "true" difference between teachers' portfolio performance, (b) a systematic difference in rater performance (a specified error influence), and (c) a "residual" error term that combines an interaction term with other unspecified sources of error. The generalizability studies found that the unspecified error term was high (between 22% and 80%), and between 25% and 50% of the variance was attributable to actual differences between teachers' portfolios. Error due to raters has generally been small (between 3% and 10%, depending on the dimension and subject). Generalizability coefficients have ranged between 0.31 and 0.76 depending on the portfolio dimension, year, and subject matter analyzed. In a 2010 generalizability and decision study, for the current NTES correction process of one rater and one occasion, the G-Index was 0.73, and the Phi-Index was 0.43. These indexes show that generalizability of the NTES scoring process is adequate regarding the ordering of the final scores (relative decisions), but not when decisions are based on the individual-level score (absolute decisions). Although the current system already contemplates double-rating of 20% of portfolios, double-rating of 100% of portfolios should be explored as an option, to ensure higher generalizability coefficients (for more details, see Taut et al., 2012).

Along with the studies mentioned above, every year the NTES technical staff carry out qualitative studies of the scoring process such as rater training, motivation of teachers that apply for being a rater, use of scoring materials, or raters' cognitive processing while scoring a portfolio (García, Torres, & Leyton, 2013).

Peer interview development

The peer interview is conducted by a classroom teacher who works in the same educational level (e.g. pre-school, elementary or secondary) as the evaluated teacher but not in the same school. The interview usually takes about 50 minutes and is conducted at the evaluated teacher's school. Each peer interviewer is usually asked to conduct various interviews (about 12 on average), depending on his or her geographic location and the number of evaluatees in his or her vicinity. A peer interviewer is selected to be peer interviewer only if he or she is not scheduled to be evaluated that year.

The peer interview consists of between six and eight questions, which are the same for all grade levels and subject areas. Every year at least three pilot studies inform the development of the questions and rubrics for the peer interview, involving about one hundred classroom teachers, as well as technical experts. They evaluate the relevance and clarity of the questions, the focus of possible answers, and the rubrics, and the final study involves a pilot application with about 50 teachers from different schools; these interviews are taped, transcribed, and analyzed before the actual implementation.

The type of questions included in the peer interview has changed slightly over the years from more specific and concrete questions to questions that demand a more complex and comprehensive

answer, in which the teacher must demonstrate his or her pedagogical knowledge more directly. For example, a question might ask what is the purpose of implementing intermediate assessments (as opposed to diagnostic and summative assessments) or what factors the teacher considers when selecting teaching tools and resources. Sometimes the teacher has to describe how he or she faces a specific situation that challenges his or her teaching skills, for example, having to adjust the time assigned for different learning activities.

Peer interviewer selection and training

Peer interviewers are selected based on the following prerequisites, as specified by law: a) They must be municipal teachers with at least four years of teaching experience, b) they must not have been subject to disciplinary actions, and c) if they have already been evaluated, they need to have obtained a competent or outstanding performance assessment. Preference is given to teachers who have won special recognition such as accreditation by the teaching excellence program. About N=1, 300 teachers are pre-selected and participate in the two-day training session. The training covers a) how to conduct the interview, b) how to take comprehensive notes, and c) how to score the answers based on a detailed scoring rubric, as well as d) ethical considerations in applying the peer interview. Training includes role-playing, note-taking, and scoring exercises. About 8% of the worst performing peer interviewers (based on observations by the trainer and the interviewers' scores on the exercises) are asked to leave the process.

Strengths and weaknesses of peer interview development and scoring

Peer interviews are not based on direct evidence but on verbal declarations of the evaluatees, which may or may not reflect teachers' actual practice. In fact, a disadvantage of the peer interview is that it provides room for cheating since knowing the scoring rubric of the interview clearly opens the door to providing the desired answers. Therefore, confidentiality regarding the interview questions and the rubric is a delicate and essential requirement, involving many hundreds of peer evaluators each year. The OECD report about the NTES criticizes that the peer interview does not involve interaction and feedback for the evaluatee and considers that it should contribute more directly to professional development (Santiago et al., 2013). For example, direct observation by the peer evaluator would provide more valid evidence and could also provide the opportunity for peer discussion and feedback based on actual classroom practice, but this is logistically complex to implement since it would mean that the peer evaluator would have to coordinate classroom hours with the evaluatees.

A positive byproduct of the peer interview is the professional development that interview training represents for the evaluators. Thousands of teachers have received training in assessment and evaluation issues, including use of rubrics and the standards for good teaching. In fact, one of the strengths of the Chilean approach to teacher evaluation is the strong involvement of classroom teachers as evaluators, both as portfolio raters and as peer evaluators (OECD, 2013b). Indeed, an important group of teachers has thus obtained deeper knowledge of the evaluation, increased the legitimacy they award the system and decreased their resistance to being evaluated. In a 2012 survey for peer evaluators, 71% of them reported that they had an improved opinion of the NTES after their peer evaluator experience.

Supervisor assessment development

This instrument consists of a questionnaire that is filled out by the principal as well as the head of the school's technical-pedagogical unit for each teacher evaluated in a school. Each questionnaire accounts for 5% of teachers' final score. Supervisors must rate each teacher's performance on a number of indicators, on a scale from 1 ("unsatisfactory") to 4 ("outstanding").

The mean score for this instrument has been stable and high over the years, ranging between "competent" and "outstanding" performance. In 2009, the scale was changed to an 8-point scale; however, this change was reversed in 2010, since the change did not result in lower mean ratings overall, despite a successful pilot-test that had indicated lower score inflation. In 2010 another innovation was introduced to attempt to control score inflation: supervisors are shown a rubric describing each performance level for each indicator. In addition, each time the supervisor assigns an "outstanding" performance level he or she must justify this decision in writing. If no justification is presented, the performance rating is automatically lowered to "competent." This change in the instrument resulted in considerably lower mean scores (by 0.3 points on a scale from 1 to 4) and a much more limited use of the "outstanding" category; in fact, the mode changed from "outstanding" to "competent" (see Fig.3), and this trend has been stable since 2010.

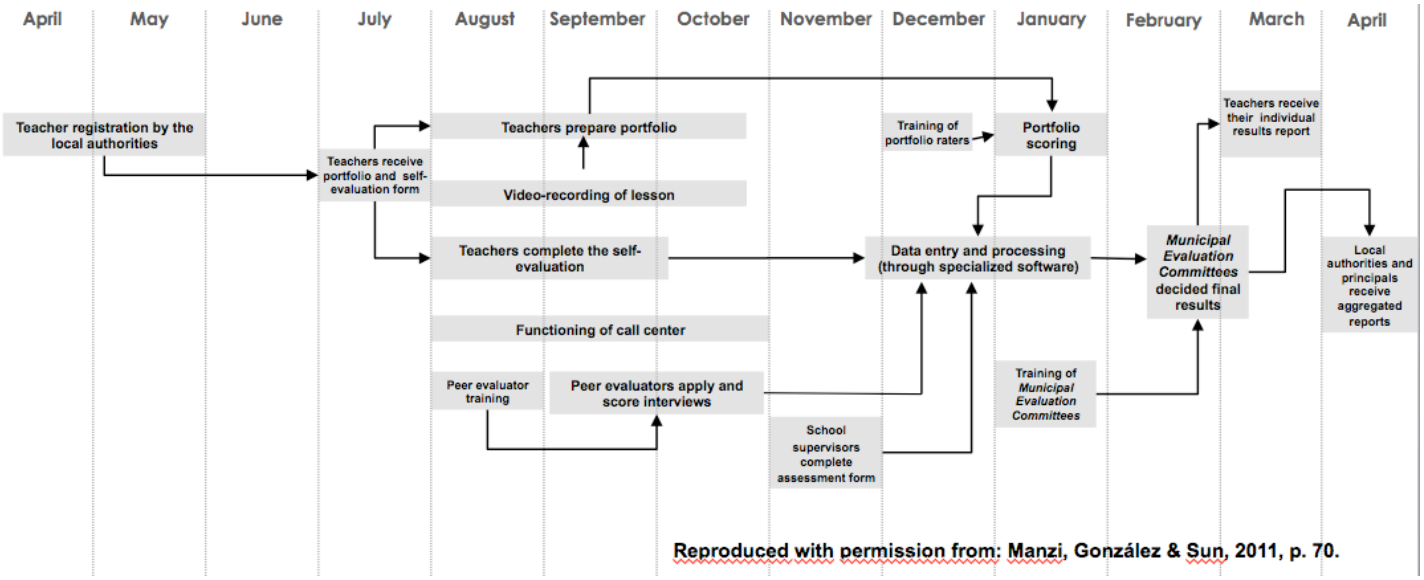

Figure 3. NTES implementation process and timeline.

Different procedures have been used to develop the questionnaire, including interviews and focus groups with school leaders, as well as pilot applications. Since 2010 the NTES staff has provided short training sessions for principals about the topic of personnel evaluation in general, and the instrument in particular. Between 2010 and 2013, N=2, 385 principals attended these meetings. In addition, since 2010 the NTES website includes a separate information section targeted to school leaders.

Strengths and weaknesses of the supervisor assessment

One weakness of the supervisor instrument has always been the inflated scores. This seems to be aggravated by the fact that the school leaders are not held accountable for their personnel evaluation practices. However, as mentioned before, recent innovations in instrument design have succeeded in controlling the problem of score inflation somewhat. The evaluation of school leaders themselves is another pending topic, which in turn might help improve their teacher evaluation practice. Another issue for discussion is that the instrument counts for only 10% of the final score. Recent evidence suggests that the evaluation might benefit from increasing the supervisor assessments' weight in the overall score (Alvarado et al., 2012; Santiago et al., 2013). If principals

played a more prominent role in the NTES they might be forced to use the evaluation more as a developmental process and ensure a more direct link with teacher professional development.

On the positive side, principals have in fact gained more control over teacher evaluation since they have been recently been allowed to fire 5% of their teaching staff who received a basic or unsatisfactory NTES result. Stakeholders agree that school leaders should have a voice in evaluating their teachers (Ministry of Education, 2012).

Self-evaluation

This instrument consists of a structured questionnaire with statements taken from the teaching standards (FGT). Teachers assess their performance on a scale from 1 ("unsatisfactory") to 4 ("outstanding"), using a simple quantitative rubric: three behavioral indicators are given and the level of performance should be assigned according to the number of them that are usually present in teachers' practice. Except for 2011, the instrument has always included some form of open-ended question in which teachers are asked to provide additional support for their self-evaluation. Since 2012 this open-ended section has been linked to the use of the outstanding performance category: every time a teacher uses this performance level he or she must describe the behaviors or practices that support such an evaluation. Likewise, in order to foster the use of the self-evaluation as an input for performance and feedback conversations between teachers and their supervisor, since 2012 both the self-evaluation and the supervisor questionnaires assess the same indicators. This is meant to ensure a common base for such conversations.

Strengths and weaknesses of the self-assessment

If self-evaluation is considered as part of the final score of a high-stakes performance assessment, this automatically leads to score inflation. Currently, the mean score for this instrument is almost 4 ("outstanding") (see Table 4). The need for justifying "outstanding" ratings in writing did not show the effect that was obtained for the supervisor assessment. The existing evidence leaves no doubt that in the current form the NTES self-evaluation is not a valid instrument. Given its weight in the final result and the consequences of the evaluation, score inflation is not surprising. In order to preserve the real value of a self-evaluation exercise, this instrument should not have weight in the final result but should be given other uses, for instance, as an input for the municipal evaluation commission (Manzi, González, & Sun, 2011a). However, since instrument weights are prescribed by law, any change would require a change in the legal framework.

Contextual considerations

Three instruments record contextual factors that may have an impact on the teacher's performance during the year of his or her evaluation. The self-assessment provides a space for the teacher to comment on these contextual issues. The supervisor questionnaire contains a similar space (see Fig.4). Finally, the peer interview offers the evaluated teacher the opportunity to comment on any issue that might have arisen since the self-assessment was handed in.

Contextual Considerations

The Framework for Good Teaching acknowledges the complexity of the teaching and learning process and the varied cultural contexts in which they are developed. Teachers do not teach in a void; they teach to specific students and in specific contexts whose conditions and particularities must be considered when their teaching is evaluated.

If there are situations or conditions of the school or the students may have affected positively or negatively the professional performance of the teacher you are evaluating in this report, please describe them briefly in the following box.

Figure 4. Contextual considerations as included in 2013 supervisor questionnaire.

These context-related comments (recorded on separate sheets of paper) are archived in the municipality in order to be reviewed by the municipal evaluation commissions when they meet to ratify or modify their teachers' final evaluation category. Although the commissions have the right to modify the final performance category, their training encourages them to do so only when they have clear and strong arguments. In fact, the commissions modify only about between 4% and 5% of the scores each year (generally 3–4% upward and 0–1% downward) (Leal & Santelices, 2010; Sun, Calderón, Valerio & Torres, 2011).

How these contextual considerations should be interpreted is not well defined, and seems a politically and ethically complex issue. The training sessions for the municipal evaluation commissions only contain a few general guidelines. For example, if a teacher blames his or her performance on the social background of poor students, this would generally not be a valid contextual consideration. However, if there were infrastructure or material shortages that made work especially difficult, or if he or she was known in the community to work well with peers and parents, then this would be something the commission could consider to raise the performance category.

In general, the exposure to the evaluation helps build evaluation capacity at the local level. When the commissions meet, all members have to understand how the system works, the standards the system is based on, what each instrument contains, as well as the quality of the evidence the instruments provide. However, the composition of the commission could be reconsidered, for example, by including the school principals of the evaluated teachers as important stakeholders.

NTES-Related Validation Research

Concern about the validity and reliability of the NTES instruments is reflected in the processes underlying NTES development and implementation, as described above, but external validation studies that are not part of the regular assessment development process have also contributed valuable information regarding the technical quality of the NTES. This more externally implemented validation agenda includes construct and consequential validity studies that have focused on validating the *overall evaluation result*. To study the relationship between NTES performance and teaching practices assessed with a mix of alternative instruments and indicators, a 2006 study examined whether the NTES identified (and, consequently, rewarded or punished) the "right" teachers as high- or low-performing (Santelices, & Taut, 2011). Researchers selected a sample of 58 teachers who were evaluated by the NTES in 2005 as either "outstanding" (N=32) or "unsatisfactory" (N=26). In-depth teaching performance data were collected on both groups: three

classroom observations, expert assessments of an alternative portfolio, teachers' participation in a subject and pedagogical knowledge test, and student testing at the beginning and the end of the school year. Analyses included correlations, group comparisons, as well as hierarchical linear modeling of longitudinal student data. The study found that "outstanding" teachers showed significantly better performance than the "unsatisfactory" teachers on half of the study's indicators, and showed positive but not significant differences on the remaining indicators. Especially strong differences (with medium to large effect sizes) related to time on task during lessons, lesson structure, student behavior, and the quality of classroom assessment materials. There were also significant correlations between the results the teachers in the study sample obtained and the results these same teachers had obtained one year earlier on the NTES, with the exception of the NTES self-evaluation. The study provides solid evidence of the validity of the NTES to differentiate among extreme groups of teachers based on their pedagogical practices in the classroom. A study comparing teaching practices of competent versus basic teachers is currently under way (Taut, Jimenez, & Manzi, 2011).

NTES validity evidence also includes longitudinal data assessing student learning for outstanding versus unsatisfactory teachers (N=1, 044 students, N=40 teachers). These data were analyzed using hierarchical linear modeling (HLM) and showed that teacher performance on the NTES is a significant predictor of student achievement at the end of the school year, controlling for student achievement at the beginning of the year (Santelices, & Taut, 2011). Another recent study linking teacher evaluation results with longitudinal student achievement data (based on 2004–2006 SIMCE panel data) showed that teachers' value-added indices in mathematics, and to lesser extent in language arts, correlated significantly with their NTES results, especially with their portfolio results (Taut et al., 2013).

Several other studies have analyzed the relationship between teachers' NTES scores and student achievement as measured by SIMCE. These studies either matched individual student-level achievement at a specific point in time with teachers' NTES performance, or they used students' SIMCE results as well as teachers' NTES results aggregated at the school level. However, a limitation of these studies is that the student-level data used do not reflect students' learning gains over time (Alvarado et al., 2012; Bravo, Falck, González, Manzi, & Peirano, 2008; Manzi, Strasser, San Martín, & Contreras, 2008; Ministry of Education 2008, 2009, 2010, 2011, 2012). The results tend to support the positive relationship between teacher performance on the NTES and the SIMCE achievement of the students these teachers worked with.

Furthermore, a comprehensive study regarding the *consequential validity* of the NTES applied a mix of methods. First of all, descriptive analyses of existing databases reflecting the effects of the NTES were done, for example, describing relevant aspects of the professional development plans that are mandatory for teachers with a basic or unsatisfactory result (Cortés, Taut, Santelices, & Lagos, 2011), or teachers' participation in the incentive program AVDI (Asignación Variable por Desempeño Individual), and teachers' job trajectories (and likelihood to leave their jobs) depending on their evaluation results (Taut, Santelices, & Valencia, 2010). Second, qualitative research at the municipal, school, and individual teacher levels, via personal interviews and focus group discussions, examined the intended and unintended consequences of the NTES for stakeholders at the different levels of the educational system (Cortés et al., 2011; Taut, Santelices, Araya et al., 2010; Taut, Santelices, Araya et al., 2011; Tornero, & Taut 2010; Santelices, Taut, Araya & Manzi, in press). These studies indicate that the NTES achieved some of its intended consequences while falling short on others. Table 6 provides a summary assessment, based on the empirical evidence, for each intended use or consequence. For example, the NTES portfolio development process generally fostered peer collaboration, while the overall evaluation process had more mixed effects on schools'

work climate (Taut, Santelices, Araya et al., 2011). There is also evidence that in some municipalities (those with stronger technical capacity) the evaluation results are used for local educational planning (Santelices et al., in press), and that in some schools (those with strong pedagogical leaders and a trusting work climate) the process and results serve an internal reflection purpose (Taut, Santelices, Araya et al., 2011).

Table 6
Summary assessment of empirical findings regarding the NTES's intended consequences

Intended use of the NTES	Initial empirical evidence
Rank order teachers depending on their teaching practice	+
Diagnose strengths and weaknesses of teachers' practice	+
Inform educational decision-making at the local level	(+)
Improve job prospects by providing incentives (via AVDI).	0
Support professional development (via PDP).	0
Invite social reinforcement of teachers showing good practice	(+)
Reinforce collaboration among peers	+

Note: + means substantial or consistent evidence available; (+) means limited or heterogeneous evidence available; 0 means no evidence available.

In addition, the consequential validity research identified multiple, important unintended consequences, both positive and negative. On the positive side, psychological and motivational support was offered to low-performing teachers by school and municipal actors. On the negative side, teachers reported work overload due to the assessment process, resistance (although in diminishing intensity), negative emotions triggered by the evaluation process and the results, and the attempt to avoid evaluation using legal means and loopholes. In summary, the researchers concluded that the NTES had mixed effects for the different stakeholders, with somewhat less favorable effects on individual teachers and more favorable effects on schools and municipalities (Taut, 2013).

Discussion and Lessons Learned from Developing and Implementing the NTES

The NTES has been implemented for 10 years. We think that the experiences gained in this program and the research conducted on it provide valuable clues for improving the NTES, as well as for designing and implementing teacher assessment programs in other contexts. In what follows we attempt to provide a critical reflection of the NTES experience based on the evidence presented above, and to share some conclusions and lessons learnt.

The NTES is an example of a standards-based assessment that does not include student achievement gains as an indicator of teaching performance. This is due not only to the lack of feasibility of a value-added teacher assessment system[6] but also reflects a conscious decision based on the multiple positive features a standards-based evaluation offers and is consistent with the recommendations of prominent experts in the field (Baker, Barton, Darling-Hammond, Haertel,

[6] Measuring teacher effectiveness based on student achievement would require an enormous expansion of the present coverage of standardized testing, as well as a substantial amount of financial and technical resources.

Ladd, Linn, & Shepard, 2010; Darling-Hammond, 2012; OECD, 2013a, 2013b; Santiago et al., 2013). Standards explicate the kind of practices and behaviors that teachers are expected to implement in their classrooms. Teachers are encouraged to become familiar with the standards and can use them to analyze and reflect about their work. The standards also serve as the basis for detailed feedback to evaluated teachers about their strengths and weaknesses. In contrast, evaluation systems largely based on student achievement might identify highly effective or highly ineffective teachers, but do not provide information about the reasons for these results, thus diminishing the value of the assessment as a professional development tool. In our view the most relevant aspect in policies for improving teaching quality is assessing and improving classroom practices, and the NTES provides direct evidence in this regard (Ravela, 2011). In the future we believe that standards about subject-specific pedagogy should be included in the NTES.

Although student test results are not included in the NTES, it is important to remember that they have been used for validating the overall NTES results as well as each of its instruments. For several years the national SIMCE reports have included data showing that the NTES results are directly associated with students results: students exposed to a larger number of teachers with good NTES performance showed better test results (Ministry of Education, 2008, 2009, 2010, 2011). Additional research studies analyzed each NTES instrument in relation to students' SIMCE scores (Alvarado et al., 2012) and applying value added methodology (Taut et al., 2013). In summary, the Chilean experience illustrates a prudent use of student test results in the context of a standards-based teacher evaluation system.

Along with the standards-based nature of the NTES, its participatory design and installation process was a key aspect in enhancing the assessment's feasibility and legitimacy. The involvement of the teacher union in negotiation process was particularly relevant in Chile – a negotiation process that took a decade to come to a consensus. In other national contexts teacher union involvement has also been identified as crucial in order to diminish teachers' resistance against the installation of evaluation policies (Santiago, & Benavides, 2009).

Another important aspect in the design of a teacher evaluation policy is the careful definition of its intended purposes, uses and effects. An assessment program's underlying theory should be clearly identified and disclosed, so that it can be monitored and evaluated over time. In the case of the NTES, only legal regulations were available to help infer these intended uses and effects. In fact, researchers embarked on an empirical delineation of NTES' program theory by interviewing representatives of all stakeholder groups involved in the installation process (Taut, Santelices, Araya et al., 2010).

Another important issue in this context is the feasibility of the Chilean attempt to achieve both high-stakes summative, as well as formative (i.e., professional development) purposes. For some, such coexistence is impractical, while others consider it acceptable, or at least politically unavoidable (Herman, & Baker, 2009). So far the Chilean experience has shown that the combination is feasible, but it involves a complex and challenging trade-off. For example, score inflation is an important concern in a high-stakes summative context. In addition, honest self-reflection becomes difficult to achieve, as the NTES self-evaluation instrument shows. The OECD review about the NTES states: "Attributing high stakes to the results of Docentemás has led the developmental function of teacher evaluation to become subsumed into the accountability aim of the system." (Santiago et al., 2013, pp.170). Indeed, the NTES emerged within the context of a growing pressure for performance-based educational accountability, particularly regarding teachers (Assael, & Pavez, 2008; Avalos, & Assael, 2006; Cox, 2003). On the other hand, NTES' formative purpose was consistent with the political goal of contributing to the professionalization of teachers and the improvement of teaching and learning in Chilean public schools (Bonifaz, 2011). While we

share the concern regarding the intended formative impact of the NTES, there is some hopeful evidence in some municipalities and schools that the NTES can inform formative uses, given certain contextual prerequisites (Sisto, Montecinos, & Ahumada, 2013; Taut, Santelices, & Manzi, 2011). This suggests that there is room for improving the developmental function of the evaluation, especially through a greater involvement of local authorities and school leaders.

The most important mechanism designed to serve the NTES' formative purpose are the Professional Development Plans (PDP), which are mandatory for basic and unsatisfactory teachers. In our view, this is one of the most essential yet weakest spots of the NTES overall design. For example, the PDPs would benefit from greater quality assurance (along with systematic impact assessment), as well as a critical look at the relationship between available resources and expected results. According to Cortés et al. (2011), there is a lack of alignment between the characteristics of successful teacher development as described by the literature, and the PDPs as they are implemented in Chile. For example, the PDPs are not integrated into teachers' daily work, not sustained over time, and not performed within a community that supports learning. The PDPs should be designed so that they are perceived as a necessary and attractive learning opportunity rather than a formal administrative and stigmatizing obligation for low-performing teachers. In fact, data show that the assessment should not only mobilize teachers who are legally forced to attend the PDPs, but all teachers in the system. Most teachers demonstrate weaknesses in their pedagogical work as assessed by the portfolio. As a case in point, N=9, 249 (84.2%) out of the N=10, 989 teachers with a competent result in 2012 attained only a basic result in the portfolio. For example, an important area in which most teachers show need for improvement is classroom assessment and student feedback.

Regarding the NTES instruments, the portfolio has proven to be the most technically robust. The research about its characteristics and results (the lowest among all four instruments), scoring process, and relation to student achievement provides empirical evidence for this claim. However, pending issues include an external study of content validity and standard-setting procedures for cut-off scores. Only recently, the NTES technical team added some elements of the Body of Work standard-setting method to define the cut-off scores that differentiate the four performance levels (Cizek & Bunch, 2007). Finally, in terms of reliability, 100% of portfolio double scoring is strongly recommended in the context of a high-stakes assessment.

Regarding the other NTES instruments, the data have consistently shown that the self-assessment has virtually no discriminatory capacity, nor reflective value in the context of a high-stakes assessment. Self-assessment should serve exclusively formative purposes and could be linked to performance-related conversations held between the teacher and the school principal. The supervisor assessment has a higher, yet still limited discriminatory capacity. Supervisors should become more engaged in the evaluation process and contribute actively to its formative purposes. Principals need to incorporate teacher evaluation and feedback in their day-to-day leadership practices and take greater advantage of the evaluation process and results for human resource management. Finally, the peer interview in its present form does not seem a particularly valuable evaluation instrument, but its value in helping to install an evaluation culture should not be underestimated. Every year about 1, 300 teachers participate in the peer evaluator training and have the opportunity to familiarize themselves with the NTES. These teachers then help disseminate a better-informed view of the assessment system. In the future, peer evaluators could play a more formative role if the interview was combined with classroom observations and post-observation feedback sessions (Santiago et al., 2013).

Another important aspect of the evaluation system is the way in which the results from the four instruments are combined to define each teacher's final performance level. In Chile, this combination is defined by law and reflects the complex, highly political negotiation process. The

evidence after 10 years of implementation clearly suggests a modification of this aspect based on the validity and reliability of each instrument. Research strongly supports the use of multiple instruments and sources for teacher assessment (Isoré, 2009; Kennedy, 2010; OECD, 2013a, 2013b) but how to weigh each instrument in the final score is less clear and depends on whether the main concern is about measurement reliability, potential for useful formative feedback, or correlation with student achievement (Bill and Melinda Gates Foundation, 2013), or whether weights reflect a political consensus, like in the Chilean case.

Equity concerns for the NTES have to do with the role of the local level in modifying or ratifying the final performance category. As pointed out before, the final decision regarding each teacher's performance category rests with the local evaluation committees, thus "returning" the evaluation to the local level and giving them the possibility to take into account relevant contextual information that otherwise could not be considered in a standardized system. However, such local involvement also introduces a level of inequality into the system. For example, we see unwarranted suspension from the evaluation process for "reasons beyond the control of the evaluated teacher." There is considerable variation in the frequency of these suspensions across municipalities, and it is difficult to prove infractions if medical or psychological reasons are given as justification. This is interesting as it shows the kind of risks in terms of equity associated with evaluation programs that rest importantly (or even exclusively) on local decisions.

Although an OECD review of educational policies in Chile criticized the cost of the NTES (OECD, 2004), in an international comparison the Chilean system is not an expensive program, for example, compared to the National Board for Professional Teaching Standards (NBPTS) teacher certification process implemented in the U.S.[7] The same system implemented in other national contexts would obviously have different costs. Challenges have to do with the considerable number of teachers evaluated each year, the amount of materials that have to be distributed to and from the municipalities, the portfolio scoring process in various scoring centers around the country, and the necessary capacities that key actors in the process have to demonstrate.

The NTES offers a lot of data on teaching performance, including videotaped lessons and portfolio evidence. There are some good examples of their use in scientific research (Cornejo, Silva, & Olivares, 2011; Galdames, Medina, San Martín, Gaete, & Valdivia, 2011; Martinic, 2011; Milicic, Rosas, Scharager, García, & Godoy, 2008; Preiss, 2009; Preiss, 2011; Preiss, Larraín & Valenzuela, 2011; Radovic & Preiss, 2010) and supporting initial and in-service teacher training.[8] Nevertheless, more intentional, systematic, and strategic research and policy actions are required to take full advantage of this information.

One interesting feature of the Chilean experience is the research related to the quality of the NTES instruments (in terms of validity, reliability and equity), as well as regarding the uses and consequences of this assessment system.[9] These research findings have helped inform some adjustments and improvements of the NTES within a relatively rigid legal framework. Along with these external studies, the technical staff in charge of program design and implementation have systematically collected information from different sources and actors, for example, surveys for

---

[7] As mentioned before, the cost of the NTES is about US$400 per teacher, which represents about one eighth of what a teacher must pay to be evaluated by NBPTS.

[8] FONDEF Project Nº D09I1063 "Generación de una videoteca de buenas prácticas docentes para la formación inicial y contínua de profesores y profesoras de Chile" [Generation of a video library of good teaching practices for initial and in-service training of Chilean teachers].

[9] FONDECYT Project No. 1080135 "Validez consecuencial de los sistemas de evaluación y asignación de excelencia pedagógica docente en Chile" [Consequential validity of the teacher evaluation system and the certification of teaching excellence system in Chile].

evaluated teachers, peer evaluators, and portfolio raters, online questionnaires for local authorities, systematization of questions received by the call center, and focus groups with school principals, among others. Almost 10 years of implementation have demonstrated unequivocally that any program, no matter how carefully designed, substantially benefits from improvements based on this kind of internally and externally implemented research.

Finally, the NTES experience should also inform new policy initiatives related to teacher accountability in Chile, particularly regarding the currently limited capacities of school principals as evaluators, and the benefits and challenges of local versus centralized high-stakes assessment systems. The largely positive evidence regarding the NTES provides arguments in favor of extending teacher performance assessment to schools that receive state subsidies but are privately managed. Many have argued that these teachers should also be subject to accountability and support mechanisms (Santiago et al., 2013). Furthermore, the NTES experience should inform the definition of a teacher career ladder that articulates elements of performance assessment, professional development, and salary progression.

References

Alvarado, M., Cabezas, N., Falck, D., & Ortega, M. E. (2012). *La Evaluación Docente y sus instrumentos: Discriminación del desempeño docente y asociación con los resultados de los estudiantes.* Retrieved from website of the United Nations Development Programme: http://www.pnud.cl/areas/ReduccionPobreza/2012/EvaluacionDocente.pdf

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

Assael, J., & Pavez, J. (2008). La Construcción e Implementación de un Sistema de Evaluación del Desempeño Docente Chileno: Principales Tensiones y Desafíos. *Revista Iberoamericana de Evaluación Educativa, 1*(2), 41-55.

Avalos, B., & Assael, J. (2006). Moving form resistance to agreement: The case of the Chilean teacher performance evaluation. *International Journal of Educational Research*, *45*(4-5), 254–266. http://dx.doi.org/10.1016/j.ijer.2007.02.004

Baker, E. L., Barton, P. E., Darling-Hammond, L., Haertel, E., Ladd, H. F., Linn, R. L., & Shepard, L. A. (august/2010). Problems with the Use of Student Test Scores to Evaluate Teachers. *EPI Briefing Paper*, *278*. Retrieved from: http://www.epi.org/publications/entry/bp278

Bill and Melinda Gates Foundation. (2011). *Learning about Teaching. Initial Findings from the Measures of Effective Teaching Project.* Retrieved from website of the Bill and Melinda Gates Foundation: http://www.gatesfoundation.org/college-ready-education/Documents/preliminary-findings-research-paper.pdf

Bill and Melinda Gates Foundation. (2012). *Gathering Feedback for Teaching. Combining High-Quality Observations with Student Surveys and Achievement Gains.* Retrieved from website of the Measures of Effective Teaching Project: http://www.metproject.org/downloads/MET_Gathering_Feedback_Practioner_Brief.pdf

Bill and Melinda Gates Foundation. (2013). *Ensuring Fair and Reliable Measures of Effective Teaching: Culminating findings from the MET Project's Three-Year Study.* Retrieved from website of the Measures of Effective Teaching Project: http://metproject.org/downloads/MET_Ensuring_Fair_and_Reliable_Measures_Practitioner_Brief.pdf

Bonifaz, R. (2011). Origen de la Evaluación Docente y su conexión con las políticas públicas en

Educación. In J. Manzi, R. González & Y. Sun (Eds.), *La Evaluación Docente en Chile* (pp. 13-31). Santiago, Chile: MIDE UC.

Bravo, D., Falck, D., González, R., Manzi, J., & Peirano, C. (2008). *La relación entre la evaluación docente y el rendimiento de los alumnos: Evidencia para el caso de Chile*. Retrieved from website of the Centro de Microdatos: http://www.microdatos.cl/docto_publicaciones/Evaluacion%20docentes_rendimiento%20escolar.pdf

Centro Nacional de Evaluación para la Educación Superior. (2000). *Estándares de Calidad para Instrumentos de Evaluación Educativa.* Ciudad de México: Author.

Cizek, G. J., & Bunch, N. B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests*. Thousand Oaks, CA: Sage.

Cohen, D. K., Raudenbush, S. W., & Ball, D. L. (2003). Resources, instruction, and research. *Educational Evaluation and Policy Analysis, 25,* 119–142. http://dx.doi.org/10.3102/01623737025002119

Cornejo, C., Silva, D., & Olivares, H. (2011). Microgénesis de la enseñanza: zoom en el modo en que los profesores presentan contenidos disciplinarios. In J. Manzi, R. González & Y. Sun (Eds.), *La Evaluación Docente en Chile* (pp. 197-199). Santiago, Chile: MIDE UC.

Cortés, F., Taut, S., Santelices, V., & Lagos, M. J. (January, 2011). *Formación continua en profesores y la experiencia de los Planes de Superación Profesional (PSP) en Chile: Fortalezas y debilidades a la luz de la evidencia internacional.* Paper presented at the second annual meeting of the Sociedad Chilena de Políticas Públicas, Santiago, Chile.

Cox, C. (2003). Las políticas educacionales de Chile en las últimas dos décadas del siglo XX. In C. Cox (Ed.), *Políticas Educacionales en el Cambio de Siglo. La Reforma del Sistema Escolar en Chile* (pp.19-114). Santiago, Chile: Editorial Universitaria.

Danielson, C. (1996). *Enhancing Professional Practice: A Framework for Teaching.* Alexandria, VA: Association for Supervision and Curriculum Development.

Darling-Hammond, L. (2012). Desarrollo de un enfoque sistémico para evaluar la docencia y fomentar una enseñanza eficaz. *Pensamiento Educativo, Revista de Investigación Educacional Latinoamericana, 49*(2), 1–20. http://dx.doi.org/10.7764/PEL.49.2.2012.1

Darling-Hammond, L., Wei, R. C., & Johnson, C. M. (2009). Teacher Preparation and Teacher Learning. A Changing Policy Landscape. In G. Sykes, B. L. Schneider, & D. N. Plank (Eds.), *Handbook of Education Policy Research* (pp. 596–612). New York, NY: Taylor & Francis.

Decree No. 192 [Ministry of Education] Reglamento sobre Evaluación Docente. Diario Oficial de la República de Chile. June 11st, 2005.

Educational Testing Service. (2002). *ETS Standards for Quality and Fairness.* Princeton, NJ: Author.

Galdames, V., Medina, L., San Martín, E., Gaete, R., & Valdivia, A. (2011). ¿Qué actividades realizan los docentes de NB1 para enseñar a leer en situación de evaluación docente? Enfoques tras las prácticas pedagógicas. In J. Manzi, R. González & Y. Sun (Eds.), *La Evaluación Docente en Chile* (pp. 200-203). Santiago, Chile: MIDE UC.

García, M., Torres, P., & Leyton, C. (2013). Representaciones cognitivas involucradas en la corrección de portafolios docentes. *Pensamiento Educativo. Revista de Investigación Educacional Latinoamericana. 50*(1), 21-39. http://dx.doi.org/10.7764/PEL.50.1.2013.3

Hattie, J. (2009). *Visible Learning: A Synthesis of Over 800 Meta-Analyses Relating to Achievement*. London & New York: Routledge.

Heneman III, H. G., Milanowski, A., Kimball, S. M., & Odden, A. (may/2006). Standards-Based Teacher Evaluation as a Foundation for Knowledge-and Skill-Based Pay. *CPRE Policy Brief,*

RB-45. Retrieved from website of the Consortium for Policy Research in Education: http://www.cpre.org/index.php?option=com_content&task=view&id=68&Itemid=73

Herman, J., & Baker, E. (2009). Assessment Policy: Making Sense of the Babel. In J. Skyes, B.L. Scneider, D.N. Plank, & T. G. Ford. (Eds.), *Handbook of Education Policy Research* (pp. 176-190). New York, NY: Routledge.

Isoré, M. (2009). Teacher Evaluation: Current Practices in OECD Countries and a Literature Review. *OECD Education Working Papers*, *23*. OECD Publishing. doi: 10.1787/223283631428. Retrieved from website of the Organisation for Economic Co-operation and Development Library: http://dx.doi.org/10.1787/223283631428

Joint Committee on Standards for Educational Evaluation. (2009). *The Personnel Evaluation Standards: How to Assess Systems for Evaluating Educators* (2nd Ed.). Corwin Press.

Kennedy, M. (2010). *Teacher Assessment and the Quest for Teacher Quality: A Handbook*. San Francisco, CA: Jossey-Bass.

Law 19.961 [Ministry of Education] Sobre Evaluación Docente. Diario Oficial de la República de Chile. Santiago, Chile. August 14th, 2004.

Law 20.501 [Ministry of Education] Calidad y Equidad de la Educación. Diario Oficial de la República de Chile. Santiago, Chile. February 26th, 2011.

Leal, P., & Santelices, M. V. (2010). *Análisis Decisiones tomadas por las Comisiones Comunales de Evaluación 2005-2008*. Internal Technical Report MIDE UC, Pontificia Universidad Católica de Chile, Santiago, Chile.

León, M., Manzi, J., & Paredes, R. (2008). *Calidad Docente y Rendimiento Escolar en Chile: Evaluando la Evaluación*. Unpublished manuscript, Universidad Católica de Chile, Santiago, Chile.

Manzi, J., González, R., & Sun, Y. (2011a). Conclusiones. In J. Manzi, R. González & Y. Sun (Eds.), *La Evaluación Docente en Chile* (pp. 241-248). Santiago, Chile: MIDE UC.

Manzi, J., González, R., & Sun, Y. (2011b). *La Evaluación Docente en Chile*. Santiago, Chile: MIDE UC.

Manzi, J., Strasser, K., San Martin, E., & Contreras, D. (2008). *Quality of Education in Chile: Final Report of the Interamerican Development Bank Project*. Washington, DC: BID. Retrieved from website of the Inter American Development Bank: http://www.iadb.org/res/laresnetwork/files/pr300finaldraft.pdf

Martinez, J. F. (October, 2012). *Evaluación Docente y Validez: Aspectos Conceptuales y Metodológicos*. Paper presented at the First Latin American Congress of Educational Measurement and Evaluation (COLMEE), Santiago, Chile.

Martinic, S. (2011). Uso del tiempo e interacciones profesores - alumnos en la sala de clases. In J. Manzi, R. González, & Y. Sun (Eds.), *La Evaluación Docente en Chile* (pp. 204-208). Santiago, Chile: MIDE UC.

Milicic, N., Rosas, R., Scharager, J., García, M., & Godoy, C. (2008). Diseño, Construcción & Evaluación de una Pauta de Observación de Videos para Evaluar Calidad del Desempeño Docente. *Psykhe. 17*(2), 79-90. http://dx.doi.org/10.4067/S0718-22282008000200007

Ministry of Education. (2004). *El Marco para la Buena Enseñanza*. Santiago, Chile.

Ministry of Education. (2008). *Resultados Nacionales SIMCE 2007*. Retrieved from website of the Agency for Quality Education: http://www.agenciaeducacion.cl/wp-content/files_mf/informenacionalderesultadossimce20072.5m.pdf

Ministry of Education. (2009). *Resultados Nacionales SIMCE 2008*. Retrieved from website of the Agency for Quality Education: http://www.agenciaeducacion.cl/wp-content/files_mf/informenacionalderesultadossimce20081.8m.pdf

Ministry of Education (2010). *Resultados SIMCE 2009*. Retrieved from website of the Agency for Quality Education: http://www.agenciaeducacion.cl/wp-content/files_mf/informenacionalderesultadossimce2009870kb.pdf

Ministry of Education (2011). *SIMCE - Resultados Nacionales SIMCE 2010*. Retrieved from website of the Agency for Quality Education: http://www.agenciaeducacion.cl/wp-content/files_mf/informenacionalderesultadossimce2010247mb.pdf

Ministry of Education. (2012). Evaluación Docente y resultados de aprendizaje: ¿Qué nos dice la evidencia? *Serie Evidencias*, 1(6). Retrieved from website of the National Teacher Evaluation System: http://www.docentemas.cl/docs/2012/EvDocente_result_aprendizajeSerie%20EvidenciasMINEDUC2012.pdf

Ministry of Education. (2013). Resultados Evaluación Docente 2012. Retrieved from website of the National Teacher Evaluation System: http://www.docentemas.cl/docs/Resultados_Evaluacion_Docente_2012.pdf

Organisation for Economic Co-operation and Development. (2004). *Reviews of National Policies for Education: Chile*. Paris, France: Author. doi: 10.1787/9789264106352-en. Retrieved from website of the Organisation for Economic Co-operation and Development Library: http://dx.doi.org/10.1787/9789264106352-en

Organisation for Economic Co-operation and Development. (2013a). *Synergies for Better Learning: An International Perspective on Evaluation and Assessment*. OECD *Reviews of Evaluation and Assessment in Education*. Paris, France: Author. doi: 10.1787/9789264190658-en. Retrieved from website of the Organisation for Economic Co-operation and Development Library: http://dx.doi.org/10.1787/9789264190658-en

Organisation for Economic Co-operation and Development. (2013b). *Teachers for the 21st Century: Using Evaluation to Improve Teaching*. Paris, France: Author. Retrieved from website of the Organisation for Economic Co-operation and Development Library: http://www.oecd.org/site/eduistp13/TS2013%20Background%20Report.pdf

Orellana, R., & Merino, J. (2013). Predictores asociados a variaciones en puntajes SIMCE en la Región del Biobio. *Cultura-Hombre-Sociedad. 23*(1), 37-54.

Preiss, D. (2009). The Chilean instructional pattern for the teaching of language: A video-survey study based on a national program for the assessment of teaching. *Learning & Individual Differences. 19*(1), 1-11. http://dx.doi.org/10.1016/j.lindif.2008.08.004

Preiss, D. (2011). Patrones instruccionales en Chile: La evidencia de la Evaluación Docente. In J. Manzi, R. González, & Y. Sun (Eds.), *La Evaluación Docente en Chile* (pp. 209-210). Santiago, Chile: MIDE UC.

Preiss, D., Larraín, A., & Valenzuela, S. (2011). Discurso y Pensamiento en el Aula Matemática Chilena. *Psykhe. 20*(2), 131-146. http://dx.doi.org/10.4067/S0718-22282011000200011

Radovic, D., & Preiss, D. (2010). Patrones de Discurso Observados en el Aula de Matemática de Segundo Ciclo Básico en Chile. *Psykhe. 19*(2), 65-79. http://dx.doi.org/10.4067/S0718-22282010000200007

Ravela, P. (2011). Aportes para pensar las políticas de evaluación docente. In J. Manzi, R. González, & Y. Sun (Eds.), *La Evaluación Docente en Chile* (pp. 222-230). Santiago, Chile: MIDE UC.

Rose, M. (2011). The Mismeasure of Teaching and Learning: How Contemporary School Reform Fails the Test. *Dissent*, *58*(2), 32–38. http://dx.doi.org/10.1353/dss.2011.0042

Santelices, M. V., & Taut, S. (2011). Convergent validity evidence regarding the validity of the Chilean standards-based teacher evaluation system. *Assessment in Education: Principles, Policy & Practice, 18*(1), 73. http://dx.doi.org/10.1080/0969594X.2011.534948

Santelices, M, Taut, S., Araya, C., & Manzi, J. (in press). Consecuencias a Nivel Local de un Sistema de Evaluación de Profesores: El Caso de Chile. *Revista Estudios Pedagógicos*.

Santiago, P., & Benavides, F. (2009). *Teacher Evaluation: A Conceptual Framework and Examples of Country Practices*. Paris, France: OECD Publishing.

Santiago, P., Benavides, F., Danielson, Ch., Goe, L., & Nusche, D. (2013). *Teacher Evaluation in Chile 2013. OECD Reviews of Evaluation and Assessment in Education.* Paris, France: OECD Publishing. Retrieved from website of the Organisation for Economic Co-operation and Development Library: http://dx.doi.org/10.1787/9789264172616-en

Sisto, V., Montecinos, C., & Ahumada, L. (2013). Disputas de significado e identidad: la construcción local del trabajo docente en el contexto de las Políticas de Evaluación e Incentivo al Desempeño en Chile. *Universitas Psychologica. 12*(1), 173-184.

Sun, Y., Calderón, P., Valerio, N. & Torres, P. (2011a). La Implementación de la Evaluación Docente. In J. Manzi, R. González, & Y. Sun (Eds.), *La Evaluación Docente en Chile* (pp. 65-89). Santiago, Chile: MIDE UC.

Sun, Y., Correa, M., Zapata, A., & Carrasco, D. (2011b). Resultados: Qué dice la Evaluación Docente acerca de la enseñanza en Chile. In J. Manzi, R. González, & Y. Sun (Eds.), *La Evaluación Docente en Chile* (pp. 91-135). Santiago, Chile: MIDE UC.

Taut, S. (August, 2013). *High-stakes Teacher Evaluation in Chile: Professionalizing Teachers in a Perfect Market?*. Paper presented at the Bi-annual conference of the European Association for Research on Learning and Instruction (EARLI), Munich, Germany.

Taut, S., Jiménez, D., & Manzi, J. (2011). *Validation of the Chilean National Teacher Evaluation System Using Student Learning progress and in-depth Examinations of Teaching.* Proposal approved for funding by the FONDECYT National Research Funding Competition, 2012 (No. 1120441). CONICYT-FONDECYT, Santiago, Chile.

Taut, S., Santelices, M. V., Araya, C., & Manzi, J. (2010). Theory underlying a national teacher evaluation program. *Evaluation and Program Planning, 33*(4), 477–486. http://dx.doi.org/10.1016/j.evalprogplan.2010.01.002

Taut, S., Santelices, M. V., Araya, C., & Manzi, J. (April, 2011). *Effects and uses of the national teacher evaluation system in Chilean elementary schools*. Paper presented at the annual conference of the American Education Research Association, New Orleans, LA.

Taut, S., Santelices, M. V., & Manzi, J. (2011). Estudios de Validez de la Evaluación Docente. In J. Manzi, R. González, & Y. Sun (Eds.), *La Evaluación Docente en Chile* (pp. 157-175). Santiago, Chile: MIDE UC.

Taut, S., Santelices, M. V., & Stecher, B. (2012). Validation of a national teacher assessment and improvement system. *Educational Assessment Journal, 17*(4)., 163–199. http://dx.doi.org/10.1080/10627197.2012.735913

Taut, S., Santelices, M. V., & Valencia, E. (2010). *Resultado de re-evaluaciones y situación laboral de los docentes evaluados por el Sistema de Evaluación de Desempeño Docente entre 2003 y 2008*. MIDE UC, Pontificia Universidad Católica de Chile, Santiago, Chile.

Taut, S., Valencia, E., Santelices, M. V., Palacios, D., Jiménez, D., & Manzi, J. (August, 2013). *Relationship between multiple measures of teaching quality and student learning: Evidence from Chile.* Paper presented at the biannual conference of the European Association for Research on Learning and Instruction (EARLI), Munich, Germany.

Tornero, B., & Taut, S. (2010). A mandatory, high-stakes national teacher evaluation system: Perceptions and attributions of teachers who actively refuse to participate. *Studies in Educational Evaluation, 36,* 132–142. http://dx.doi.org/10.1016/j.stueduc.2011.02.002

About the Authors

Sandy Taut
School of Psychology / MIDE UC, Pontificia Universidad Católica de Chile
staut@uc.cl ; staut@ucla.edu
Sandy Taut obtained her professional degree in Psychology from the University of Cologne, Germany and her Ph.D. in Education from the University of California Los Angeles (UCLA). She is an assistant professor at the School of Psychology at Pontificia Universidad Católica de Chile and an associate researcher at the Measurement Center MIDE UC.

Yulan Sun
MIDE UC, Pontificia Universidad Católica de Chile
ysun@uc.cl
Yulan Sun has been the manager of the Docentemás project, the team at MIDE UC advising the Ministry of Education on the Chilean National Teacher Evaluation Program, since 2003. Before joining MIDE UC she worked for the Chilean Ministry of Education. She holds a Masters degree in Psychology from Pontificia Universidad Católica de Chile.

education policy analysis archives
Volume 22 Number 71 July 21st, 2014 ISSN 1068-2341