Validation of the Chilean national teacher evaluation system[1]

Sandy Taut
Psychology Department, MIDE UC
Pontificia Universidad Católica de Chile

María Verónica Santelices
School of Education
Pontificia Universidad Católica de Chile

Brian Stecher
RAND Corporation, Santa Monica, USA

Abstract

Chile has a national teacher evaluation system (NTES) that is standards-based, uses multiple instruments and is intended to serve both formative and summative purposes. For the past six years the authors have performed validation research on NTES using a variety of methods and data sources. This paper describes our validation research agenda, the results of major validation studies, an integration of the existing evidence, and offers the authors' preliminary judgment about NTES' validity. The paper also offers a critical reflection regarding the decisions taken while driving the long and winding validation road, and the pending issues to be addressed.

[1] We thank partial funding of this work by the Chilean Government grant FONDECYT 1080135.

About the authors

Sandy Taut (staut@uc.cl) is an educational researcher and adjunct professor at the Psychology Department, Pontificia Universidad Católica de Chile (PUC), where she heads the research unit of the Measurement Center MIDE UC. Her research has focused on educational measurement, consequences of educational accountability, and teacher assessment. Sandy Taut obtained her professional degree in Psychology from the University of Cologne, Germany, and a Ph.D. in Education from the University of California, Los Angeles (UCLA).

María Verónica Santelices (vsanteli@uc.cl) is an assistant professor at the School of Education of the Pontificia Universidad Católica de Chile (PUC). Her interests include educational measurement, evaluation, standardized testing, access to higher education and educational policy. She worked as an analyst and researcher in different national and U.S. educational institutions. She received her doctorate in Education and a Master in Public Policy from the University of California at Berkeley, and a professional title in Public Administration from PUC.

Brian Stecher is a Senior Social Scientist and the Associate Director of RAND Education. Dr. Stecher's research focuses on measuring educational quality and evaluating education reforms, with a particular emphasis on assessment and accountability systems. At RAND he has directed prominent national and state evaluations, currently a seven-year study of the Gates Foundation Intensive Partnership for Effective Teaching. Stecher has served on expert panels relating to standards, assessments, and accountability for the National Academies, and is member of the National Research Council's Board on Testing and Assessment. He received his Ph.D. in Education from the University of California, Los Angeles (UCLA).

Introduction

Improving teacher effectiveness has become a major focus of educational reform policy in the U.S. However, existing efforts to measure teacher effectiveness and use the information as a tool to improve student performance have not been thoroughly evaluated. The experiences of educators in Chile, who are implementing and validating a national teacher evaluation system (NTES), can serve as a model for researchers and policymakers in the U.S.

Both the states and the federal government are undertaking efforts to improve teacher effectiveness. These programs include expanding the way teachers are evaluated (to provide better data about teacher effectiveness), changing the way teachers are managed (to more effectively support them, more equitably place them with students whose needs are the greatest, more efficiently compensate them in line with their value to the system), and changing teachers' career paths (to more thoughtfully use them in different roles befitting their expertise). A few examples illustrate the efforts currently underway. The federal government supports and encourages these policies through Teacher Incentive Fund (see http://www2.ed.gov/programs/teacherincentive/index.html) and the Race to the Top, which has as one pillar rewarding effective teaching (see http://www2.ed.gov/programs/racetothetop/phase3-resources.html; Center on Education Policy, 2011). The Teacher Incentive Fund is funding districts to develop new evaluation and compensation systems for teachers (and in Pittsburgh for school leaders). Race to the Top has evaluating teacher effectiveness in raising student achievement as one of its four priorities. Late in 2011 the Department of Education established rules allowing states to request waivers from certain NCLB mandates if, among other actions, they implemented systems to evaluate teacher effectiveness. The Bill and Melinda Gates Foundation has funded the Intensive Partnership for Empowering Effective Teachers (IPS), a $300 million effort that is trying to transform teacher evaluation and human capital management in four large school districts[2].

Given the importance assigned to these teacher evaluation and improvement policies and the resources devoted to them, it is appropriate that they be thoroughly evaluated. The Standards for Educational and Psychological Testing (hereafter short "the Standards"), as well as a number of researchers, call for a comprehensive validation agenda regarding large-scale educational assessments, especially those with high-stakes consequences (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999; Linn, 2006; Kane, 2006).

However, comprehensive validation of assessment and improvement systems is not an easy task. These systems are complex entities that include standards describing quality performance, strategies for measuring performance, classification of effectiveness, and differential consequences, ranging from incentives for good performance to supports for improvement to sanctions (even dismissal) for poor performance. Evaluating the quality of such a system is an

[2] The College Ready Promise, one of the four sites, is actually a consortium of charter management organizations, but we refer to all four sites as districts for the sake of simplicity.

equally complex process (Wolming & Wikström, 2010; Gaertner & Pant, 2011).

The scope of the effort may explain why comprehensive validation efforts regarding large-scale assessment systems have not been documented extensively in the literature. There are a few exceptions, including the National Board for Professional Teaching Standards (NBPTS) certification of teaching excellence, as well as the Performance Assessment for California Teachers (PACT) (National Research Council, 2008; Ingvarson & Hattie, 2008; Pecheone & Chung, 2007). A few examples of similar validation efforts can be found in the context of student assessment, when states have to present validation evidence to the U.S. Department of Education in response to No Child Left Behind (Schafer, Wang & Wang, 2009; Linn, 2009).

Given the paucity of examples of comprehensive validation of large-scale assessment and improvement systems, this paper attempts to describe how such a comprehensive validation agenda was implemented in the case of a large-scale teacher performance assessment system in Chile, the national teacher evaluation system (NTES). The authors are based in the research department of the university measurement center responsible for developing and implementing NTES. For the past six years we have been performing validation research on NTES using a variety of methods and data sources.

The paper begins with a description of the Chilean teacher assessment system. Then we review the literature on assessment validation and we present our specific validation framework and research agenda. A brief discussion of methods follows, highlighting the kinds of approaches that we have used to address different questions over the past six years. The heart of the paper reviews the evidence assembled to address the various validity questions, ranging from content through consequences. We close with a preliminary judgment about NTES' validity and a discussion of the implications of our efforts for the validation of large-scale teacher assessment and improvement systems in other contexts.

Description of the Chilean national teacher evaluation system (NTES)

Introduction of the NTES

The Chilean national teacher evaluation system (NTES) was introduced amid some controversy by the Ministry of Education in 2003, and since 2005 has been mandatory for teachers in municipal schools nation-wide. Prior to the NTES, the teacher evaluation systems in Chile was locally implemented, related to seniority more than to performance, subjective in nature, had little to do with a teacher's classroom performance, and the results often had little concrete consequences for teachers (Avalos & Assael, 2006). Thus, the NTES constituted a completely new approach to developing the country's teacher workforce. Not surprisingly, the political process of introducing the NTES was characterized by difficult negotiations with the most important political stakeholders and resistance on the part of a considerable segment of teachers. But in 2002 a committee consisting of teacher union representatives, representatives of the local municipal authorities, and Ministry of Education personnel arrived at a consensus

to conduct teacher evaluation in roughly its current form. The Teacher Union consulted its members and these approved the compromise (Avalos & Assael, 2006).

Purposes and consequences of NTES

The NTES was designed to improve the quality of teaching in Chile by explicitly judging each public school teacher's effectiveness and providing this information to teachers, schools and municipalities to support improvement efforts and contribute to personnel decisions. There is a tension between the evaluation system's formative, improvement-oriented purpose and the summative uses of its results for accountability. On the summative side, those teachers receiving an "unsatisfactory" result have to be re-evaluated the following year. If they receive another "unsatisfactory" result they are subject to loss of employment in the municipality where they worked. Teachers showing "basic" performance – about a third of evaluated teachers so far – must be re-evaluated two years later and they get three chances to improve their performance to the expected performance level of "competent" before being subject to termination. [3] Teachers who are high-performing ("competent" or "outstanding") are eligible to take a subject knowledge test, and depending on the test results, they receive a salary bonus for up to four years until their re-evaluation.

At the same time, there is a formative purpose to NTES in at least two ways: (1) Individual reports are sent to each evaluated teacher, indicating strengths and weaknesses on a number of dimensions, defining the expected performance for each dimensions and contrasting this information with the observed performance. (2) Low performing teachers are offered professional development in their respective municipalities. These professional development opportunities are supposed to address the weaknesses NTES has diagnosed for each teacher. The professional development activities are monitored via an online tool used by municipalities, implementers and teachers.

In addition, every year NTES sends reports to each school communicating the results of its evaluated teachers, which implies that the developers of the system intended the results to be used at school level. Finally, NTES is expected to inform educational policy not only at local but also at national level by providing overall information on the quality of teachers in the country's public schools. All the above-mentioned purposes and uses presuppose that NTES correctly distinguishes high-performing from low-performing teachers.

As we noted, the creation of NTES was contentious. Its goals and purposes were not always explicit, and different interest groups emphasized different aspects of the system. Consequently, an initial part of our research agenda was to study the intended operation and effects of NTES. We analyzed relevant policy and legal documents and interviewed program stakeholders from

[3] These consequences are in effect since 2011 (Law No. 20.501). Prior to 2011, the consequences for low performance were less severe: In case of unsatisfactory performance, the teacher had to be reevaluated the following year, but had two more chances to improve his or her performance, instead of just one. Basic teachers had to undergo reevaluation only after four years, instead of after two years, and there were no punitive consequences attached to repeated basic performance.

the Education Ministry, the teacher union and the municipalities association who had been part of the initial negotiations, as well as implementers of the program. Based on the documents and interviews we developed lists of intended effects and uses, and we assessed their importance for different stakeholder groups. We identified the following six major effects; each was important to at least two stakeholder groups: Maintaining good practices by triggering internal reinforcement of diagnosed strengths and offering social reinforcement of good teaching practices; triggering change of weak teaching practices and building the capacity of teachers with shortcomings; diagnosing the quality of teaching practices as a basis for management decisions; informing the selection of new teachers and the exit of unsatisfactory teachers; providing a base for peer collaboration on good practice; improving teachers' job prospects by offering monetary incentives. These six intended effects will be taken up again when we discuss the results of our study of NTES' consequential validity. In the long term, interviewees agreed that the evaluation policy should result in an improvement of students´ learning (for details see Taut, Santelices, Araya & Manzi, 2010). Thus, the stakeholders recognized that the NTES served dual purposes, providing formative data to improve teaching practice and providing summative data to support rewards and sanctions.

Standards for Effective Teaching

An important step in the introduction of the NTES was the publication of professional teaching standards in 2004, called "Marco para la Buena Enseñanza [Guidelines for Good Teaching]" (Ministry of Education, 2004). These professional teaching standards were developed on the basis of Danielson's Framework for Teaching (Danielson, 1996). The Teacher Union was involved in the development of these standards and formally approved them (Avalos & Assael, 2006). The document clusters standards into four major domains: Domain A "Preparation for Teaching", Domain B "Creating a Learning Environment", Domain C "Opportunity to Learn for All Students", and Domain D "Professional Responsibilities."

Components of the NTES

The NTES consists of four assessment components, including a portfolio assessment (comprising a written part and a videotaped lesson), a supervisor assessment, a peer interview, and a self-assessment. Brief descriptions of each are presented in the following paragraphs.

Portfolio Assessment. The portfolio asks teachers to describe lesson planning and classroom evaluation materials for a specific, pre-defined set of lessons (a teaching unit), as well as to reflect on the use of these materials in the classroom. In addition, one lesson (45 minutes) from each teacher is videotaped by an external contractor. Portfolio instructions and specific items of the scoring rubrics differ across years, while the general structure of the portfolio and the dimensions of the scoring rubric remain constant across time. The eight dimensions are: A. Planning of teaching unit; B. Analysis of teaching and learning activities; C. Quality of classroom assessment; D. Reflection based on classroom

assessment results; E. Reflection about own practice; F. Classroom learning climate; G. Structure of lesson; and H. Pedagogical interaction. The first five dimensions are evaluated by the written part of the portfolio, the last three by the video-taped lesson. Each of the eight dimensions of the portfolio rubric is operationalized by three items, so the rubric contains a total of 24 items. For example, in Dimension B the associated items cover the teacher's strategies to address the learning difficulties presented by the students, the identification of elements of a teaching unit that helped students progress in their learning, and the identification of aspects of the content that serve as barriers for such learning. In the portfolio instructions related to Dimension B, the teachers would be asked to describe the particular teaching unit, reflect on the unit in terms of pedagogical activities and contents, and analyze the learning difficulties of their students (Flotts & Abarzúa, 2011).

Supervisor Assessment. Two supervisors (generally the principal of the school and the teacher in charge of the so-called Technical Pedagogical Unit) complete an evaluation questionnaire asking about professional qualities of the evaluated teacher. Examples would be whether the teacher collaborated with colleagues and interacted appropriately with parents. Since 2010 two-hour trainings are provided to a sub-sample of municipal evaluation coordinators and school leaders. In 2010, about N=520 principals attended the training.

Peer Interview. The peer interview is performed by a teacher from a different school, who teaches the same subject and grade level as the teacher being evaluated. The interviewer follows a structured interview protocol containing questions about pedagogical knowledge and practice. Each peer evaluator candidate participates in a two-day training (about N=1400 are pre-selected, of which the 10% worst performing are asked to leave the process). This process is intended to ensure valid and reliable evaluation data and also to build evaluation capacity and contribute to the installation of an evaluation culture.

Self-Assessment. Finally, the self-assessment is a questionnaire that asks the teacher to self-evaluate aspects of professional performance and to reflect on his or her performance as a teacher.

Scoring, Scaling and Performance Levels

Each item of each instrument is responded to or scored by raters on a four-point scale: 1=unsatisfactory; 2=basic; 3=competent; 4=outstanding. In the case of the supervisor and self-assessments, the items are responded to directly on the four-point scale, while in the case of the peer interview and the portfolio scoring is done based on the evidence and by applying the rubrics that define the four points on the scale for each item. The notes taken during the peer interview are later scored by the same interviewer, while the portfolio is scored by an independent rater. Total scores for each instrument are based on calculating the mean across items. In the case of the portfolio, items are combined into dimension scores, and dimension scores into the overall score. Each item within a dimension, and each dimension in the overall portfolio, has the same weight.

Since portfolio scoring is the most complex part of the evaluation and

contributes most weight to the overall score (see below), we will focus attention on this aspect. Each year the portfolios and corresponding scoring rubrics are developed based on the following sources of information:

(a) the official teaching standards (Marco para la Buena Enseñanza);

(b) videos of teaching practice in each subject and grade level are consulted to make sure the portfolio links to actual classroom practice;

(c) a team of expert teachers pertaining to the different subjects and grade levels, as well as technical experts, perform a detailed revision of portfolio instructions and rubrics;

(d) pilot studies are implemented for each year's draft portfolios, applying the think-aloud technique both individually and in groups of teachers;

(e) pilot studies are implemented for 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;

(f) data from the scoring process of the previous year are consulted to identify problematic items.

In terms of assuring the quality of the rating process, all raters are trained during 30 hours, to get to know the scoring rubric and learn to apply it by using the same practice portfolios, while their performance is being monitored constantly. The professionals in charge of supervising the rating process also receive training (40 hours) during the month prior to the scoring period. There is also a three-day trial period when all processes are tested and raters continue practicing and being monitored. The training and dry-run scoring helps to identify those supervisors and raters who diverge from the pre-established scores and to replace them if low performance persists. Every Monday during the three-week scoring period all raters complete a group scoring session with their supervisors. Also, twenty percent (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, and divergent raters are retrained on the use of the rubric.

The scores obtained by each instrument have different weights in the final performance categorization, as defined by law (see Law 19.933, 2004; Law 19.961, 2004; Decreto N°192, 2004): the portfolio assessment contributes 60% of the final score; peer interview 20%; and supervisor and self-assessment 10% each (see http://www.docentemas.cl/).

Performance level descriptions are developed sequentially every year for each item that is scored on the four-point performance scale, starting by defining "competent" performance, followed by defining "basic", then "unsatisfactory" and lastly "outstanding" performance. Operationalized definitions of "competent" performance are extrapolations of the teaching standards, applied to each subject and grade level separately, and based on the professional judgment of the pedagogical and disciplinary experts involved in NTES.

According to NTES data of N=66,938 teachers obtaining final evaluation results between 2003 and 2010, the majority (60%) of evaluated teachers receive the performance categorization of "competent", while a third are evaluated as showing "basic" performance (30%). Only a small percentage (8%) is evaluated

as “outstanding”, and even fewer (1%) are considered “unsatisfactory”. These results show little variation each year (Sun, Correa, Zapata & Carrasco, 2011).

Modifications to the evaluation process and the final scores

Every municipal teacher goes through the evaluation process every four years. However, there are some conditions under which the NTES is not administered or the results are subject to local modification. These special circumstances were initiated as part of the negotiations to create the NTES in response to specific concerns of stakeholders. One of the key questions to be investigated is whether they are supporting the goals of the NTES or reducing its validity and fairness.

The legal regulations of NTES allow teachers to suspend the evaluation process, or to be exempted from the process for a handful of reasons. Teachers are exempted from the process if they are in the first year of their teaching career, work as peer evaluators, or –since 2007- are within three years of the legal retirement age. The evaluation process gets suspended if the teacher changed schools or classes during the year of evaluation, or had a substantial leave of absence due to illness or sabbatical leave. Suspension can also be granted for “unforeseeable circumstances”, and the interpretations of what this means rests with the local educational authorities overseeing the evaluation process at municipal level. An example of such circumstances is the 2010 earthquake, which substantially changed teaching conditions in part of the country during the 2010 school year. In a typical year without big natural disasters about 12% of teachers suspend their evaluation based on such “unforeseeable circumstances”, and this justification makes up the large majority of suspensions in the system (Leal & Santelices, 2010).

Another feature that stresses the role of local educational authorities in the NTES process are the so-called “local evaluation commissions.” These commissions are composed of the peer evaluators of the municipality and the person in charge of the NTES in the municipality. The commission receives the final evaluation results of all the teachers evaluated by NTES in any given year before they are communicated to teachers and schools. It is charged with confirming or modifying the results, taking into account local contextual information provided by the evaluated teacher and his or her supervisors. Each modification must be justified in an on-line database. So far only about 3% of final scores are modified every year (Manzi, González & Sun, 2011).

Finally, teachers can formally object the evaluation result they obtained and initiate an appeals process. This leads to a case-by-case review by the Ministry of Education in collaboration with the implementing institution.

Reporting NTES results

Reports with the results of the evaluation process are sent to each evaluated teacher, communicating his or her overall performance category based on the combined performance on all four instruments, as well as the performance levels by instrument (the supervisor and peer assessments are shown as a composite score) and for each of the eight portfolio dimensions. For each

dimension, the teacher receives a description of "competent" performance as well as the strengths and weaknesses shown in his/her particular evidence. The reports for the school include the overall performance category of each teacher who pertains to the respective school, as well as teachers' mean results obtained in the supervisor assessment on the one hand, and the portfolio and peer interview on the other hand. The report also shows teachers' mean results for each portfolio dimension. Municipal reports are similar to school reports but the level of analysis is the municipality and also includes summaries of results for each school within the municipality who had its teachers evaluated.

Related programs and policies

There is a monetary incentive program associated with the NTES. It is called Individual Performance Bonus [Asignación Variable por Desempeño Individual] (AVDI). Teachers who are found to be high-performing on the NTES ("competent" or "outstanding") are eligible for an increase in salary if they also pass a subject and pedagogical knowledge test. Their performance on the test, combined with their NTES performance, determines what will be the percentage of salary increase they receive for the next four years until their re-evaluation. The salary increase is based on the average annual salary (all public school teachers are public employees whose salaries are regulated on a national level) and varies between 5%, 15% and 25% annual increase.

Low performing teachers on the NTES, on the other hand, are subject to mandatory professional development. The so-called Professional Development Plans [Planes de Superación Profesional] (PSP) to provide this professional development are developed at the municipal level. That is, municipal authorities are charged with the design and implementation of these development opportunities. Each municipality receives federal funding based on the number of unsatisfactory and basic teachers they have in a given year in order to implement these activities. The PSPs are monitored by the Ministry using an on-line tool.

The MBE standards form the basis not only for NTES but also for a voluntary accreditation of teaching excellence program (Asignación de Excelencia Pedagógica, AEP), which was introduced in 2003. The AEP assesses teachers on the basis of a portfolio and a video-taped lesson (which inspired the NTES portfolio), as well as a subject and pedagogical knowledge test, which also serves as the AVDI test mentioned above. This program is voluntary for all teachers working in schools that receive government funding (public and private subsidized schools). AEP teachers receive an additional monthly salary a year, for up to 10 years as long as they remain teaching in the classroom. They are also offered pedagogical consulting roles giving them access to additional remuneration (Red de Maestros para Maestros). Municipal teachers must obtain competent or outstanding scores in their NTES evaluations in order to retain their AEP accreditation. We used available information about teachers who were evaluated by both programs in the NTES validation during the early years of program implementation. Later, self-selection of AEP applicants based on their NTES scores may have reduced the validity of the score comparison.

Theoretical frameworks and other literature informing our validation research

The concept of validity has evolved and expanded over time as the sophistication and use of tests has expanded. Kane (2006) provides an interesting overview of this evolution. He describes how the criterion and content models of validity, which arose from selection and achievement testing respectively, were replaced in the early 1980s by construct validation, which came to be seen as the basis for a unified model of validity (also see Messick, 1994, 1995). This formulation of validity is also reflected in the Standards for Educational and Psychological Testing (hereafter "the Standards"; American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999), which is the most widely accepted source with regard to educational validation research.

The Standards define validity as "the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests" (p. 9). The Standards advocate that validation must focus on the proposed *interpretation* and *uses* of test (or assessment) scores and that it involves a research program instead of a simple study. In fact, the Standards echo Kane (2006) in that they advocate "developing a scientifically sound validity argument to support the intended interpretation of test scores and their relevance to the proposed use" (1999, p. 9). They further indicate that "a sound validity argument integrates various strands of evidence into a coherent account of the degree to which existing evidence and theory support the intended interpretations of test scores for specific uses" (1999, p. 17). The evidence should be based on test content, response processes, internal structure, relations to other variables, and consequences of testing (pp. 11-17).

While the focus of validation must be the interpretation of scores, there is an ongoing debate among scholars about whether to include consequential aspects within a validation framework (NCME Newsletter, 2010). Kane (2006) has convincingly argued that especially if a testing program serves as an engine of reform to improve educational outcomes, then it makes sense to evaluate it as an educational program, and program evaluations include intended as well as unintended outcomes of the program being evaluated. This is important because for stakeholders to make informed decisions about the effectiveness of high-stakes tests, it is necessary that they have information about how well these tests achieve various goals and at what cost. Assuming that there are both positive and negative consequences, the stakeholders and policymakers face the task of weighing these consequences against each other (p. 56). We agree with Kane (2006), Messick (1989, 1995), Shepard (1997) and Linn (2009) on including an evaluation of both the meaning of test scores and the consequences of their use within our research program on NTES' validity.

Linn (2009) gives us further orientation as to the best methodology to investigate intended and unintended consequences of testing programs. Because there is no universal agreement on which consequences are positive or negative, he sees little purpose in separating the two when collecting these kinds

of data. He suggests applying neutral questions so respondents are free to express either positive or negative opinions. Furthermore, researchers should include qualitative, exploratory methods, in order to make sure to capture emerging unintended consequences. While it is challenging to establish a clear causal connection between an assessment program and its effects (and in fact, can hardly be expected to be unambiguous), this challenge must be confronted.

In addition to consulting the literature regarding conceptual and methodological issues of validation, we also looked at examples of other assessment systems' validation efforts to guide us in our task. Particularly helpful was the experience of the National Board for Professional Teaching Standards (Moss, 2008; National Research Council, 2008; Bond, Smith, Baker & Hattie, 2000). The literature on the NBPTS helped us think about validity evidence in a more inclusive way, not only considering studies especially (externally) performed for validation but also consulting documentation and research routinely performed by the staff of the assessment program. On the other hand, as Moss (2008) pointed out, the routine studies are mostly confirmatory in nature and do not look for competing hypotheses or interpretations, while special studies are usually performed by external researchers more independent from the assessment program itself and thus more open to disconfirming evidence. This supports the value of a validation research agenda implemented by researchers independent of the assessment development unit (also see Kane, 2006, p. 25).

The National Research Council (2008) reviewed recent content-, construct-and criterion-based validity evidence regarding the NBPTS, as well as evidence related to the reliability and fairness of the certification program. Regarding validity, they note that "certification programs generally rely on content-based validity evidence" while the NBPTS also collected construct- and criterion-based evidence (p. 110). Content-related validity evidence was collected by using expert panels that examined the appropriateness of the standards, and the congruence between the standards and the exercises and scoring rubrics. As the most extensive study on construct-based evidence the National Research Council (2008) cites a study by Bond, Smith, Baker & Hattie (2000), which investigated the classroom practice of board-certified teachers (N=31) versus unsuccessful candidates (N=34) and found better performance by board-certified teachers on all dimensions (though not always significant differences). This study also inspired one of our studies that examined in depth the classroom practice of teachers who had been evaluated by NTES as either outstanding or unsatisfactory.

The National Research Council (2008) further explains that although external criteria for validating certification tests are difficult to find and even unnecessary, they also point out that the NBPTS has a predictive component related to student achievement, and they dedicate one chapter of their review to discussing studies that investigate the relationship between teachers' NBPTS certification and student gain scores on standardized achievement tests. The general conclusion from this review is that certified teachers were overall slightly more effective in raising students' test scores than unsuccessful candidates. In

our case, robust evidence on NTES performance related to students' gain scores is still missing (only one small-size study includes such evidence).

Schafer, Wang and Wang (2009) analyzed the validity documentation included in the 2007 No Child Left Behind legislative review process of state assessment systems, in order to identify the most salient types of validity evidence presented by the states. They found that states generally provided stronger evidence based on test content, internal response processes and internal structure, while there was weaker evidence based on relations with other variables, and virtually no evidence regarding intended and unintended consequences of NCLB testing. This latter type of validity evidence is precisely what Linn (2009) calls for when discussing the validation of NCLB testing programs, arguing that validation should be seen as a comprehensive evaluative process.

Decisions we took in defining our validation research agenda

In 2005 we started investigating the validity of the NTES. Following the suggestions found in the literature, we thought to develop a comprehensive validation research agenda, i.e., an extended research program that by amassing multiple kinds of evidence would shed light on key aspects of the assessment program. We used the Standards (1999) as a starting point to structure our validation efforts. We decided to focus on validity, reliability and fairness of the NTES, and within validity we organized our work around the types of evidence delineated in the standards, i.e. evidence based on content, internal structure, relationships with other variables, and consequences. The section describing the results of our research is organized in terms of these six main ideas. In Tables 1 and 2 we present summaries of all available and still pending evidence for the NTES. While Table 1 focuses on validity evidence, Table 2 reflects other relevant aspects of the technical quality of the NTES assessment system: reliability and fairness. As mentioned above, in this paper we will report primarily the studies we were directly involved with, and those studies are shown in bold in Tables 1 and 2.

We should also note that our agenda was shaped in part by the resource constraints we confronted; we had to prioritize those studies that seemed most crucial to investigating NTES' validity, reliability and fairness (Cronbach, 1989; Anastasi, 1986; Kane, 2006). We made two important simplifying decisions when planning our research efforts. First, the proposed interpretations and uses of NTES that we take as the basis of our validation work come out of our empirical work analyzing policy documents and interviewing relevant stakeholders. As explained above, we focus on the most important purposes and uses (Herman & Baker, 2009). Second, we decided to focus on the overall assessment score because it has direct consequences for individual teachers, schools and districts, and on the portfolio instrument because it has most weight in the determination of the final score and because evidence exists that scores on the other components (supervisor-, peer-, and self-evaluation) are uniformly high (Sun, Correa, Zapata & Carrasco, 2011).

We decided to focus on the following questions regarding NTES' validity that were not addressed by the developers as part of their regular assessment development and implementation process.

Content-related evidence:

> Does the Teacher Evaluation cover the content of the Guidelines for Good Teaching (MBE)?

Evidence related to internal structure:

> Do the portfolio scores represent the different aspects (factors) of teaching performance reported by the NTES?

Evidence about relations with other variables:

> Do the highest and the lowest performing teachers (as labeled by NTES) show meaningful differences in their teaching performance when applying alternative instruments measuring similar, as well as complementary aspects of teacher performance?
>
> Are there any significant differences in the academic performance of students depending on whether their teachers attained the best or worst scores in the NTES?
>
> How does the performance of teachers assessed by NTES and by the accreditation for teaching excellence program (AEP) relate to one another?

Consequential evidence:

> Does NTES have the intended consequences, and are there important negative, unintended consequences for teachers, schools and districts?

Evidence on reliability/generalizability of portfolio scoring:

> Is portfolio scoring free from rater bias – or is there a substantial amount of variance in assessment scores due to rater effect?

Evidence on fairness of NTES' assessment process:

> Is the evaluation process fair to all evaluated teachers? Do the local evaluation commissions function based on comparable criteria, or are there important differences in how they modify teachers' final scores? Are teachers treated fairly and equally across the country regarding their right to suspend the evaluation process?

Methods

In this section we briefly describe the range of methods used across the studies we conducted to validate the NTES for its intended uses. Details can be found in the specific study reports (Santelices, Taut, Araya & Manzi, 2009; Santelices, Taut & Valencia, 2008, 2009; Leal & Santelices, 2010; Taut, Santelices, Araya & Manzi, 2011; Taut, Santelices & Valencia, 2010; Valencia & Taut, 2008) and related journal articles (Taut, Santelices, Araya & Manzi, 2010, in press; Tornero & Taut, 2010; Santelices & Taut, 2011). We applied a mix of

methods, ranging from psychometric analyses and advanced statistical analyses to qualitative studies using personal interviews and focus groups.

In 2005 one of the co-authors analyzed the alignment of the teaching standards with portfolio scoring rubrics. The analysis was done using tables mapping the standards onto the items of the rubrics after careful study of both the standards and the rubrics, and based on professional judgment (see Table 6 in Santelices & Taut, 2010).

Since 2005 we have used both exploratory and confirmatory factor analyses to study the structure of the assessment instruments, with particular emphasis on the portfolio (Valencia & Taut, 2008). While exploratory factor analysis illuminates questions regarding the number and nature of latent variables that might explain the shared variance of a matrix of correlations of portfolio items, confirmatory factor analysis offers evidence regarding a pre-established underlying portfolio structure, which in our case corresponds to the eight dimensions as eight related but distinguishable underlying constructs (Tabachnick & Fidell, 1996; Preacher & McCallum, 2003). Our exploratory factor analyses applied the Principal Axis Factoring (PAF) and Maximum Likelihood (ML) estimation methods, as well as various factor retention rules (rule of Kaiser-Gutmann, scree test, interpretability). We used rotation method Oblimin, however, reaching similar findings when applying Varimax (orthogonal) rotation. The confirmatory factor analysis was based on N= 10,350 observations, corresponding to 2010 NTES portfolios. The analyses were performed for ordered nominal data using the robust weighted least squares estimation method and a tetrachoric correlation input matrix. We performed the analyses in Mplus 5.21. In order to evaluate model fit we consulted various indices: chi-square test of model fit, comparative fit index (CFI), Tucker-Lewis fit index (TLI), Root Mean Square Error of Approximation (RMSEA). According to Brown (2006) the appropriate cut-off values are >0.95 for CFI and TLI; <0.06 for RMSEA.

In terms of relationships with other variables, we performed correlational analyses and group comparisons, as well as hierarchical linear modeling of longitudinal student data. In order to study teacher performance one year post-evaluation, we conducted classroom observations, supervised expert assessments of an alternative portfolio, and applied student testing at the beginning and at the end of the school year (Santelices & Taut, 2011; Taut & Santelices, 2007; Santelices, Taut & Valencia, 2008).

We studied consequential validity by applying a mix of methods: first of all, we descriptively analyzed existing databases reflecting the effects of NTES, for example, describing relevant aspects of the professional development program PSP, or participation in the incentive program AVDI. Second, we conducted qualitative research at municipal (local), school and individual teacher levels, via personal interviews and focus group discussions, to examine the intended and unintended consequences NTES has had for stakeholders at these different levels of the educational system (Taut, Santelices, Araya & Manzi, 2010; Tornero & Taut, 2010; Santelices, Taut, Araya & Manzi, 2009; Taut, Santelices, Araya & Manzi, 2011, in press; Taut, Santelices & Valencia, 2010; Santelices, Taut & Valencia, 2008, 2009).

We also calculated indices reflecting the inter-rater agreement, inter-rater reliability and generalizability of the portfolio scoring process (Haertel, 2006). The major advance of Generalizability theory over the classical theory definition of reliability is that it allows for a more precise specification of error in measurement (Brennan, 2001; Shavelson & Webb, 1991). In our G-studies we studied raters as a possible source of error. We used the data available from the regular portfolio scoring process of those portfolios that were double-rated by the same rater pairs. In 2010 we also conducted a quasi-experimental study where, in addition to raters, we added occasion as a facet.

Finally, evidence regarding assessment fairness came from descriptive statistics as well as content analysis of existing databases (Leal & Santelices, 2010).

Results of validation studies performed to date

This section briefly describes the specific validation studies we have performed to date to answer the research questions that guided our validation research agenda.

Evidence based on assessment content

In 2005 we studied the content validity of the NTES in terms of its alignment with the standards for good teaching described in the "Marco para la Buena Enseñanza" (MBE). The alignment study found that the portfolio covered a large majority of indicators related to Domain B "Creating a Learning Environment" and Domain C "Opportunity to Learn for All Students", and partially covered those related to Domain A "Preparation for Teaching" and Domain D "Professional Responsibilities." However, the indicators related to Domain D ("Professional Responsibilities") are assessed by other part of the NTES, the peer interview, the supervisor assessment and the self-assessment.[4] In Domain A, in particular, there is limited coverage of standards related to subject-specific pedagogy[5] and content knowledge, which also fail to be addressed by other NTES instruments. However, it is important to mention that content knowledge is a concern of NTES developers, since high-performing teachers are required to pass a disciplinary and pedagogical knowledge test before becoming eligible for the salary bonuses offered by the AVDI program. The negotiations preceding installation of NTES resulted in an explicit exclusion of measuring disciplinary knowledge as part of the *mandatory* evaluation process.

Evidence based on internal structure of NTES instruments[6]

[4] A similar alignment analysis for all NTES instruments was done by project staff in 2010 and shows good coverage of NTES as a whole, and especially of the portfolio instrument.

[5] Since 2008, a subject-specific indicator has been included in the NTES portfolio.

[6] Internal consistency using Cronbach's Alpha is routinely checked by program staff. Between 2005 and 2010, the indices ranged between 0.74 and 0.81 for the written part of the portfolio, between 0.70 and 0.79 for the video-taped lesson, between 0.96 and 0.99 for the supervisor assessment, and between 0.78 and 0.87 for the peer interview.

As mentioned previously, the scoring rubric for the NTES portfolio contains 24 indicators grouped in 8 dimensions. Since 2005 we routinely performed exploratory factor analysis of every year´s NTES portfolio data and in 2010 have also conducted confirmatory factor analysis. We have also conducted exploratory factor analysis of the other NTES instruments using data from 2005 and 2009.

Results have varied somewhat over the years but in general the exploratory factor analysis identified either five or six factors for the entire portfolio, including the written part and the videotaped lesson (see Table 3 for 2009 results). Together these factors explain between 30% and 39% of the variance in scores. Usually, three or four of the factors are associated with abilities tested in the written part of the portfolio and the other factors are associated with the videotaped lesson. The factors associated to the video-taped lesson change from year to year, but in some years have neatly recreated the underlying theoretical dimensions. The factors associated to the written portfolio have been more stable over time and we have called them (a) reflecting on pedagogical decisions; (b) designing classroom assessment materials; and (c) lesson planning. These empirical factors combine similar tasks that are repeated across dimensions, for example, the asking the teacher to reflect on the work included in the portfolio (reflect on the planning aspect as well as the assessment aspect, for example). In some years the data neatly replicated the classroom assessment dimension.

Overall, the factor structure we identified using exploratory factor analysis resembles the NTES portfolio *tasks* more than the scoring rubric *dimensions*. 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, according to the CFI, TLI and RMSEA indices (see Table 4).

As a whole, we think these results partially validate the structure of the portfolios and provide suggestions for improving future scoring and reporting. For example, it would be possible to supply feedback to teachers about their performance on the indicators that involve completing teaching tasks such as lesson planning and classroom assessment in comparison to those that entail reflection. This could be done in addition to the feedback currently provided based on the 5 dimensions of the written part of the portfolio.

The factor analyses exploring the factor structure of the other three NTES instruments (analyzing 2005 and 2009 NTES data), including the self-assessment questionnaire, supervisor assessment questionnaire, and peer assessment indicate that each of these instruments is one-dimensional, that is, they primarily depend on one latent variable. Each of them reveals that teachers, supervisors, and peers based their answers on a global assessment of teacher performance.

Evidence based on relationships with other variables

We conducted several analyses of the relationship between NTES results and other variables measuring similar constructs, and the results generally support the validity of NTES scores. The first study compared the pedagogical

practices of teachers receiving high and low scores from NTES (Santelices & Taut, 2011; Taut & Santelices, 2007). Another set of studies related teacher performance with student achievement as measured by standardized tests (SIMCE)[7]. The third study explored the relationship between teachers' NTES results and their scores in the Pedagogical Excellence Certification program [Asignación de Excelencia Pedagógica, AEP], which also involves a portfolio based on the Guidelines for Good Teaching [Marco para la Buena Enseñanza, MBE] (Santelices, Taut & Valencia, 2008). The results are summarized in the following sections.

Pedagogical practices of teachers with high and low performance in NTES. During 2006 we conducted a validity study that examined whether NTES identifies – and consequently rewards or punishes – the "right" teachers as high- or low-performing (Santelices & Taut, 2011; Taut & Santelices, 2007). We selected a sample of 58 teachers who were evaluated by NTES in 2005 as either "outstanding" (N=32) or "unsatisfactory" (N=26). We collected in-depth teaching performance data on both groups. The performance evidence included the following:

- Students´ gain scores measured by a standardized, curriculum-based achievement test administered at the beginning and at the end of the school year.
- Observations by trained research staff (blind to the NTES performance of the participating teachers) of three 90-minute lessons taught by each teacher during the school year. The observation checklist evaluated the following elements: use of time (time on task), lesson structure, stimulation of critical thinking, presence of conceptual errors, student behavior, adaptability to students' prior knowledge and questions, and pedagogical flexibility.
- Expert assessment of a set of teaching materials for a two-week curricular unit, which included planning, teaching and assessment activities, along with samples of student work. Teachers were also requested to answer a questionnaire on school context and teacher reflection. Experts were classroom teachers who had experience working with rubrics and who were especially trained and monitored (including 100% double rating).
- Teacher scores on a test of disciplinary and pedagogical knowledge (the AEP test).

We found that "outstanding" teachers had significantly better outcomes than the "unsatisfactory" teacher on half of the performance indicators, and showed positive but not significant differences on the remaining indicators. We found especially strong and practically significant differences related to time on task during lessons, lesson structure, student behavior, and classroom assessment materials (see Table 5). We also found significant correlations

[7] SIMCE is a curriculum-based standardized test that is designed and administered by the Ministry of Education to all Chilean students in 4th, 8th and 10th grade in order to monitor student learning.

between the results the teachers in our sample obtained in our study and the results these same teachers had obtained one year earlier in NTES, with the exception of the NTES self-evaluation questionnaire. The study provides solid evidence for the validity of the NTES to differentiate among extreme groups of teachers based on their pedagogical practices in the classroom.

Relationship between NTES and student achievement. Another type of evidence associated with the validity of the NTES involves longitudinal data assessing students´ learning for both outstanding and unsatisfactory teachers (N=1044 students of N=40 teachers). These data were analyzed using hierarchical linear modeling (HLM) and showed that teacher performance in 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).

Several studies have analyzed the relationship between teachers' NTES scores and student achievement as measured by SIMCE (Bravo, Falck, González, Manzi, & Peirano, 2008; Manzi, Strasser, San Martin, & Contreras, 2007; Ministry of Education 2008, 2009). Their results tend to support the positive relationship between teacher performance in NTES and SIMCE achievement of the students these teachers worked with. These studies either matched individual student-level achievement at a specific point in time with their teachers' NTES performance or students' SIMCE results aggregated at the school-level with the NTES results from a group of teachers pertaining to the same school. However, a limitation of these studies is that the student-level data they use do not reflect students' learning gains over time.

Convergence between NTES and the Certification of Teaching Excellence program. We compared the performance of teachers who were rated on both the NTES and the AEP (Santelices, Taut & Valencia, 2008). Between 2002 and 2006 we have evidence from both programs for 739 teachers. We found that the great majority (93%) of teachers being certified with teaching excellence come from high performance levels, as indicated by their NTES scores ("competent": 67%; "outstanding": 26%). Likewise, teachers who receive good results in NTES are more likely to apply to, and win, AEP than are teachers from the two lower performance categories. When comparing performance by instruments, we observe a moderate positive correlation between performance on the NTES portfolio and the AEP portfolio (0.33), as well as between NTES final score and AEP portfolio score (0.36). We find correlations below 0.3 between scores on the other three NTES instruments and the AEP portfolio, as well as between NTES instruments and the AEP subject and pedagogical knowledge test.

These results may be affected by temporal differences in the application to the programs (one or two years apart) and by differences in the nature of participating in them: while the NTES is mandatory and offers economic incentives that need an additional application and assessment process (for example, the AVDI test is not part of NTES), the participation in AEP is voluntary and carries a degree of social recognition that may motivate teachers to apply,

independent from the direct benefit of the salary bonus. We conclude that there is moderate convergent evidence regarding NTES' validity when considering the AEP assessment as criterion.

Consequential validity

We examined empirically whether NTES' intended consequences, as identified by our study of legal documents and stakeholder perceptions (see details above), were in fact observed, and what were unintended consequences. We chose two methodological approaches: (a) analyses of existing data bases related to consequences at individual teacher level, and (b) interviews and group discussions with teachers, school leaders, and municipal educational authorities. During the interviews and group discussions we set out asking general, non-leading questions regarding the programs' consequences first, and only as a second step did we probe more specific intended effects and uses. We interviewed N=19 municipal actors in 10 purposively sampled municipalities; conducted interviews with N=57 school leaders from 30 public elementary schools from those same 10 municipalities; implemented N=9 focus group discussions with a total of N=46 teachers; and interviewed N=10 teachers who received an "unsatisfactory" performance rating, as well as N=9 teachers who had at least once actively refused to participate in NTES.

Our empirical findings indicate that the assessment program achieves some of its intended consequences while falling short on others (Santelices, Taut, Araya & Manzi, 2009; Taut, Santelices, Araya & Manzi, 2011, in press; Santelices, Taut & Valencia, 2008, 2009; Taut, Santelices & Valencia, 2010). The following brief summaries address both the individual and systemic consequences:

Maintaining good practices by triggering internal reinforcement of diagnosed strengths and offering social reinforcement of good teaching practices. Based on focus group and interview evidence we found that teachers perceived recognition to be insufficient. School leaders and school district authorities report both formal and informal recognition practices, but these are inconsistent across time and informal practices are more prevalent. Also, by far the largest effect and the one that is present across levels and sub-groups of interviewees, was the experiences of negative reactions (disapproval, envy) of significant others (peers, superiors) to the assessment results teachers obtained.

Triggering change of weak teaching practices and building the capacity of teachers with shortcomings. Its formative nature is a major feature of NTES and one where it has only partially met its promise. Teachers and municipal stakeholders told us that the assessment process itself, especially engaging in developing the NTES portfolio, led teachers to revises their teaching practice.

On the other hand, professional development was not effective according to many teachers (Santelices, Taut & Valencia, 2009; Cortés, Taut, Santelices & Lagos, 2011). Professional development is mandatory for teachers receiving an "unsatisfactory" or "basic" performance level. Each municipality (school district)

receives a fixed amount of money for each low-performing teacher working in their district, which is to be spent on these teachers' professional development, as organized by the municipality. Professional development at municipal level is planed based on the shortcomings diagnosed by NTES but only roughly half of all teachers who are obligated to attend actually do so in practice, and due to its generally short duration and ineffective format the professional development is of heterogeneous and generally limited quality and impact.

We analyzed data from the 2007 to 2009 PSP online database reflecting teacher participation rates, course contents and types of delivery, as well as teacher satisfaction with this initiative. We found that over the three-year period only 41% of teachers obligated to participate actually did so. In 2008, professional development most commonly happened in the form of workshops or seminars (76%), followed by working individually with a mentor (17%), and only 3% consisting of classroom observations and feedback. The PSP most often address "classroom assessment" and "lesson planning." Teachers generally show high levels of satisfaction with the PSP (although never more than 30% answer the corresponding satisfaction questionnaire). However, 15% of respondents say they would not participate again. In terms of potential consequences, an analysis combining data from the 2009 NTES and PSP programs indicates that the unsatisfactory teachers who participated in PSP activities in 2009 (N = 64) and underwent obligatory re-evaluation had average scores above those of unsatisfactory teachers who did not take part in PSP (N = 12) (for more details, see Cortés et al., 2011). These differences were statistically significant.

Diagnosing the quality of teaching practices as a basis for management decisions. Based on municipal and school leaders' self-reports, we can say that NTES does inform educational management decisions, particularly at municipal level (Santelices et al., 2009; Taut et al., in press). All municipalities reported at least some use of NTES results for educational planning, such as assignment of teachers to schools based on their NTES performance. Similar responses were provided (albeit to lesser extent) at school level. At school level, three-thirds of schools reported that NTES serves them as external validation of their achievements, and more than half mentioned using the results as a diagnosis of the quality of their teaching, which led some of these schools to implement internal reflection and evaluation processes intended to improve teaching.

Informing the selection of new teachers and the exit of unsatisfactory teachers. This is a sub-effect of the one previously mentioned and one where NTES has generally not inspired the intended use of its results. Municipalities reported not using NTES as a tool for the selection or exit of their teachers, also because of the legal restrictions that exist in this regard. However, we performed analyses of public school teachers' job trajectories, using the Ministry's teacher salary databases for 2007-2009, and NTES databases for 2003-2008 and found that teachers who have received at least once an unsatisfactory performance rating are three times as likely to leave the public teaching force than those who

have never received such a result (32% versus 11%, see Table 6) (for details see Taut, Santelices & Valencia, 2010). This suggests that unsatisfactory teachers leave the municipal system before being evaluated as unsatisfactory for a third time. The results from this study also showed that high-performing teachers were more likely to secure administrative positions.

Providing a base for peer collaboration on good practice. This intended effect is one that was reported across all levels and stakeholder groups. Virtually all teachers and school leaders said that teachers collaborated on the elaboration of the evaluation instruments, and in many schools the assessment results are commented on and trigger shared reflection among teachers. Municipal actors also observed peer collaboration in eight out of ten municipalities.

Improving teachers' job prospects by offering monetary incentives. According to the large majority of teachers we interviewed, the AVDI incentive program associated with NTES has not had an important effect in terms of teachers' job commitment, satisfaction or career prospects (Santelices, Taut & Valencia, 2008, Table 7). We analyzed data from 2004 to 2006 regarding the participation of teachers in the incentives program associated with NTES. In order to qualify for a salary bonus, teachers have to pass a subject knowledge test corresponding to the subject matters they teach. If they pass it with a performance level of "outstanding", they receive a 25% annual salary bonus. If they are qualified as "competent" in the test, they get a 15% increase. Since 2006, even teachers who receive a performance level of "basic" in the test receive a small bonus of 5% annually. All these bonuses are paid until the teacher faces re-evaluation after 4 years. We found that of the total number of "outstanding" and "competent" teachers enabled to participate in AVDI from 2004 to 2006, only 46% in fact took the test. In total, over the three years, about half (54%) of those taking the test earned some kind of salary bonus. It is important to point out that of these teachers, on average over the three years, only 23% received a 15% salary bonus, and a total of 32 teachers received the 25% bonus.

In focus groups, teachers complained that the incentive was too low, the barriers of achieving it too high, and little information available. The perception of AVDI by municipal and school actors was somewhat more positive, in that the majority said that it served as an effective teacher incentive policy.

Unintended consequences. In addition, we identified multiple, important unintended consequences, both positive and negative. On the positive side, we found the psychological and motivational support that has been offered to low-performing teachers by school and municipal actors (including by way of the professional development). On the negative side, teachers reported a number of unintended effects: work overload due to the assessment process, resistance (although in diminishing intensity), negative emotions triggered by both the evaluation process and the results, and the attempt to avoid evaluation using legal means and loopholes.

We investigated in more depth why some teachers openly refused to participate in the NTES (Tornero & Taut, 2011). We asked teachers how they explained their own refusal to participate and on the basis of their responses we created an explanatory model for their behavior. The methodology included nine in-depth interviews analyzed according to Grounded Theory procedures (Strauss & Corbin, 1991). Our findings indicated that several factors influenced the active rejection of the NTES by this group of teachers, including strong negative emotions generated by NTES (such as fear of getting a poor result), cultural aspects of the teaching profession in Chile, as well as negative perceptions of NTES regarding its legitimacy (evaluation criteria and instruments), its compulsory nature, and the lack of information about the evaluation system.

Conclusions about consequences of NTES. From a researcher perspective we conclude that NTES has had mixed effects for the different stakeholders, with less favorable effects on teachers and more favorable effects on schools and municipalities. Without doubt the formative, professional development aspect of NTES, as well as the associated incentives program, both need to be strengthened. Of course, others might interpret the overall impact of these results differently.

Reliability and generalizability evidence

We have examined rater performance as a possible source of measurement error in the NTES portfolio in 2005, 2008, 2009 and 2010 through different types of analyses, which we report in this section.

Our findings from 2005 show that the mean inter-rater reliability (IRR) for the written part of the portfolio was R = 0.61; the mean IRR for the video-taped lesson was also R = 0.61. These indices are below what is generally regarded an acceptable IRR (R = 0.8). While Dimensions 2 to 7 obtain IRR indices between 0.61 and 0.70, Dimensions 1 and 8 stood out with especially low indices of 0.51 and 0.52, respectively.

Between 2005 and 2009 we conducted G-studies with the purpose of examining the percentage of variability in the obtained scores that was attributable to: (a) “true” difference between teacher portfolio performance (denoted “Teachers” below), (b) systematic difference in rater performance (a specified error influence, denoted as “Raters”), and (c) a “residual” error term that combines an interaction term with other unspecified sources of error (an unspecified error influence, denoted as “TxR, e”).

The G-studies have found TxR,e to be high (between 22% and 80%), and between 25% and 50% of the variance attributable to actual differences between teachers’ portfolios. Variance due to rater effect has generally been small (between 3% and 10%, depending on dimension and subject). Generalizability coefficients have ranged between 0.31 and 0.76 depending on the portfolio dimension, year and subject matter analyzed. Not all centers scored the same subject matters or grades therefore it was not possible to include these variables in the G-study maintaining a completely crossed design.

In 2010 we designed and implemented a G-study independent of the actual portfolio scoring process in which we included rater and scoring occasion (2 times, seven days apart) as facets in a completely crossed design; also, we compared the generalizability of the rater facet in two different scoring schemes: (1) each rater scored the complete written module of a single teacher (scoring by teacher) and (2) each rater scored dimension 1 in all of the portfolios assigned to him/her, then he/she continued with dimension 2, and so on (scoring by dimension). The results indicated that real differences between portfolios explained most of the variance of the written module's final score (48%), followed by differences attributable to raters (31%) and error (17%). The study showed that the scoring occasion did not explain a significant percentage of the variance in the final score. Results were better when raters scored by teacher instead of by dimension. Finally, for the current correction process of one rater and one occasion the G-Index is 0.73 and the Phi-Index is 0.43. These indexes show that generalizability of the NTES scoring process is adequate with respect to the ordering of the final scores (relative decisions), but that the raters' performance does not reach optimal levels when decisions are based on the individual level score (absolute decisions).

Additionally, in 2005 and 2009 more detailed analyses have highlighted large differences in rater performance between individual raters or rater pairs and by supervisor. This indicates that good rater performance is possible and does occur in the current system, but is not yet consistently shown by all raters. More obvious rater characteristics like age, title, institution where title was obtained, and job experience do not give a clear indication of rater quality but there are some patterns that suggest that raters' teaching experience and the number of hours they work at schools may be of importance.

In summary, reliability and generalizability of the portfolio ratings are below what would be expected of a high-stakes assessment system. Although some of the recommendations discussed in Myford and Engelhard (2001) in the context of misfitting raters in the National Board for Professional Teaching Standards (NBPTS) assessment system have already been implemented in the NTES (e.g., training raters on portfolios known to be heterogeneous in order to evaluate their use of the scoring scale; as well as having detailed information about rater performance presented in charts or tables, collected and reviewed in *real time* so supervisors can train individual scorers early on in the process) we think more could be done. Possible modifications include progressing towards the double scoring of all portfolios (not just 20%), as well as simplifying and reducing the number of dimensions evaluated and exploring a holistic evaluation rubric.

Evidence regarding fairness of the assessment process

We considered two aspects of fairness related to specific procedures that are part of the administration of NTES: exemption and suspension policies and unequal application of decisions by the local evaluation commissions.

Exemption and suspension from the evaluation process. We examined the consistency of exemption and suspension decisions made and recorded by the

local educational authority (Leal & Santelices, 2010a). In this regard, NTES' laws and regulations allow for some degree of interpretation and subjective judgment. Some reasons for exemption are explicitly stated, including teachers who have less than a year of experience, teachers who are serving as a peer evaluator that same year, and teachers being within two years of retiring. However, judgment is particularly important when suspending teachers from the process, especially when invoking "extraordinary reasons." Recording the exemption and suspension decisions is mandatory, but the recording of the specific situation behind "extraordinary reasons" is not. However, 90% of municipalities do provide such information. We examined these data from the 2005 to 2008 national evaluation processes. Between 2005 and 2008, between 20% and 30% of teachers who were supposed to undergo NTES were exempted or suspended from the evaluation. Most of them suspended the process (between 13% and 20%) and "extraordinary reasons" explained between 83% and 93% of all suspensions. Physical and mental health issues in particular were the most frequent reasons mentioned. These patterns were consistently observed across municipalities.

Considering contextual information in the final performance category. The local evaluation commissions have the prerogative to ratify or modify the final assessment category of a teacher. They can modify the final rating if they consider that there is contextual information that should be considered when evaluating a given teacher's performance. Reasons for modifying the final performance category need to be input into a monitoring software, and a special note to the teacher needs to justify any modification decision. We examined data from the 2005 to 2008 national evaluation processes and analyzed the reasons mentioned for these modifications in a representative sample of municipalities (Leal & Santelices, 2010b). We observed that local evaluation commissions generally maintained the final performance category (95% to 96% depending on the year) and that most modifications actually increased the final performance category (67% and 86% depending on the year). This increase is most frequently observed for teachers who had originally been assessed as basic and whose category was increased to competent by the local evaluation commission. This can be explained by the incentives available to teachers in the latter performance category (e.g., the AVDI incentive program). The available information was uninformative as to what type of contextual information had actually been considered by the local evaluation commissions when making their modification decisions.

Discussion

This paper described issues related to developing a validation research agenda for a large-scale teacher evaluation system and presented the validity evidence we have accumulated to date for judging the validity of this teacher evaluation system. We also presented the general framework used to conceptualize validity, which is consistent with the guidelines disseminated in the Standards, i.e., we gathered multiple types of evidence related to validity,

including consequences (AERA, APA & NCME, 1999). In this section we combine the sources of evidence that we examined during our investigation and attempt to come to a final conclusion about the validity of the NTES.

We believe that the relationship between NTES results and other variables collected to date support the validity of the NTES final category. The NTES process reveals real differences between high and low-performing teachers. The study of pedagogical practices showed substantial differences (measured by effect size) between the performance of "unsatisfactory" and "outstanding" teachers, in terms of relevant aspects of their pedagogical work in the classroom. In addition, the study found that three of the four instruments used in the NTES (especially the portfolio) have a moderate association with the instruments used in the research study (only the NTES self-assessment is unrelated). Our results comparing teachers evaluated by two similar programs (NTES and AEP), both based on the Guidelines for Good Teaching [MBE] and including portfolios, provide additional positive validity evidence. Another important criterion for establishing the validity of the NTES is the performance of students taught by the evaluated teachers. The study on pedagogical practices, which used longitudinal student data and hierarchical linear modeling, showed that students taught by high-performing teachers tended to attain better results than their peers taught by low-performing teachers. Although of small scale, the study showed that teachers' evaluation result was a significant predictor of students' post-test score, while controlling for students' pre-test score.

Other studies showed positive correlations between teachers' NTES performance and status measures of student achievement but it is important to complement these data with additional studies based on longitudinal measurements of student learning, thus establishing a more direct relationship between student achievement and teacher performance. Another, very important piece of evidence that still needs study is the validity of adjacent categories distinguished by the NTES, for example, "basic" versus "competent" performance.

Analyses showed an appropriate coverage of the contents of the Guidelines for Good Teaching [MBE] by NTES instruments and scoring rubrics. A validation of this aspect by external educational experts remains pending.

Factor analyses suggest revising somewhat the internal structure of the portfolio, and, accordingly, of scoring rubrics and reports. The portfolio structure could be simplified, reducing the number of dimensions and aligning them with the underlying teaching tasks consistently identified in the factor analyses.

Our studies of the undesired effect of the rater on NTES portfolio scores shows that this effect is generally low, although heterogeneous among different subjects and portfolio dimensions. At the same time, reliability indexes are comparable to those reported from other teacher evaluation systems (see National Research Council, 2008). Our recommendation is to increase the percentage of portfolios that are double-rated, ideally to a hundred percent. Our analyses indicated that generalizability indexes could be improved substantially in this way. In addition and since most of our internal structure and generalizability analyses dealt with the portfolio, future research should examine

the internal structure and generalizability of the self-, peer, and supervisor assessments.

Our investigation of the intended and unintended consequences of the NTES shows that there is sufficient evidence of some intended consequences identified in the program theory, namely, promoting the social recognition of good teaching practices, as well as promoting a change in weak teaching practices (in the sense that the evaluation process itself stimulates teachers to review and update their practices, particularly by becoming familiar with the standards [MBE]). Also we found sufficient evidence that the NTES provides a diagnosis of teacher quality for educational decision-making at municipal and school levels, and that the NTES process promotes peer collaboration which may foster good teaching practices. We found mixed or weak evidence regarding the following intended consequences: maintaining good practices through internal reinforcement of individual strengths; improving the practices of low-performing teachers (via PSP); and improving teachers' work prospects by granting economic incentives associated with good individual performance (via AVDI). In addition, our participants reported important unintended consequences, both positive and negative.

The results from analyses of existing databases suggest that the incentives policy associated to NTES (AVDI) is being underutilized, which could be changed, at least in part, by an increase in the amount of the incentives. Also, more information about the eligibility criteria and the amount of the bonus needs to be available for teachers.

In addition, our studies regarding the design and implementation of the professional development (PSP) associated with NTES indicate the need for modifications in order to more effectively strengthen teaching practices and fulfilling the formative promise of NTES. Currently, the PSPs face a strong negative reaction from low performing teachers whose attendance is mandatory during non-paid hours. The literature on professional development suggests longer periods of training, with more generous funding and an implementation that is more integrated into teachers´ daily activities (see AERA, 2005). A stricter quality control of the training providers is also needed.

Although our study shows that the NTES results are not being used extensively for hiring and firing teachers by municipality and school actors, we would expect this to change in the future since the new Law No. 20.501, in effect since 2011, allows principals to dismiss 5% of the school staff based on NTES performance as well as other criteria. The new Law also includes more severe consequences for basic and unsatisfactory performance.

"Suspension for extraordinary reasons" occurs in a significant proportion of teachers who are called to be evaluated each year. Most often the reasons are stated as problems of physical and mental health. On the other hand, we observed that local evaluation commissions generally maintained the NTES final performance category as generated based on the evaluation instruments, and that most modifications actually increased the final performance category of basic teachers. The available information was uninformative as to what type of

contextual information had actually been considered by the local evaluation commissions when making their modification decisions.

Given all our work on validating NTES so far, Koretz' (2009) words ring true, "validity is often presented as a dichotomy: a conclusion is either valid or not. Unfortunately, the situation is generally murkier than this. Validity is a continuum (…) Rather, some inferences are better supported than others, but because the evidence bearing on this point is usually limited, we have to hedge our bets." (p. 219). Therefore, the way in which the results about NTES' validity are judged is a matter of perspectives and values (Herman & Baker, 2009). While some may have a negative opinion merely due to the presence of one or two negative results, others may regard these as "expectable" and come to a more positive conclusion overall. Still others may insist on the need to complete the pending information with new studies before passing conclusive judgment.

In spite of all these precautions, we conclude that the initial evidence supports a positive general assessment of NTES' validity. In our opinion, the most relevant and complex evidence for such a judgment is the relationship between NTES and other ways to measure effective teaching, along with NTES' important positive consequences for all main stakeholders – despite a more mixed picture coming from teachers, especially during the first years of evaluation implementation. However, some aspects of the evaluation must be revised, which is not surprising or unexpected given the complexity of measuring teacher performance (Berliner, 2005; Ingvarson & Rowe, 2008). In our eyes, the most important improvements to be made in terms of test development refer to the revision of the internal structure of the portfolio and the double scoring of all portfolios.

Pending issues regarding the validation of NTES

Bearing in mind Cronbach's idea that validation is a "long, even interminable" process (1989, p. 151), we are aware that there will always be better evidence to gather and more studies to conduct. Although we feel we have made significant progress in the study of NTES' validity there are still some pending issues that should be addressed in the future. Among those issues we consider the most prominent:

First, in terms of content validity we lack an expert judgment of the assessment's alignment with the professional standards (Guidelines for Good Teaching); such alignment has so far only been investigated by one of the co-authors, and by NTES program staff.

Second, we need to investigate the appropriateness of the differentiation between basic and competent teacher performance since these are the categories that distinguish expected from below-acceptable performance, with attached positive versus negative consequences. So far we have only examined the validity of the extreme performance categories (unsatisfactory versus outstanding).

Third, validation specifically related to the instruments other than the portfolio is lacking. For example, the peer interview has a weight of 20% in the

final score and should be validated specifically.

Finally, we have yet to examine the relationship between the teacher quality assessment by NTES and other teacher quality measures such as those based on value-added indexes using longitudinal student achievement data.

References

AERA (American Educational Research Association) (2005). Teaching Teachers: Professional Development to Improve Student Achievement. *Research Points*, *3*(1), 1-4.

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

Anastasi, A. (1986). Evolving concepts of test validation. Annual Review of Psychology, 37, 1-15.

Avalos, B., & Assael, J. (2006). Moving from resistance to agreement: The case of the Chilean teacher performance evaluation. *International Journal of Educational Research*, *45*(4-5), 254-266. doi:16/j.ijer.2007.02.004

Berliner, D. C. (2005). The near impossibility of testing for teacher quality. *Journal of Teacher Education, 56*(3), 205 –213. doi:10.1177/0022487105275904

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

Brennan, R. L. (2001). *Generalizability theory*. New York: Springer.

Brennan, R. L. (2006). Perspectives on the evolution and future of educational measurement. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp.1-16.). Westport, CT: American Council on Education/Praeger.

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

Brown, T. (2006). Confirmatory Factor Analysis for Applied Research. New York: Guilford Press.

Center on Education Policy (2011). More to Do, But Less Capacity to Do It: States' progress in implementing the recovery act reforms. Washington, D.C.: Author. (Downloaded from http://cep-dc.org on November 22, 2011).

Cronbach, L. (1989). Construct validation after thirty years. In Linn, R. (ed.), Intelligence: Measurement, theory and public policy, pp. 147-171. Urbana: University of Illinois Press.

Cortés, F., Taut, S., Santelices, V.& Lagos, M.J. (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 presentado en el segundo encuentro anual de la Asociación Chilena de Políticas Públicas, 19 enero 2011, Santiago, Chile.

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

Flotts, P. & Abarzúa, A. (2011). El modelo de evaluación y los instrumentos [The

evaluation model and the evaluation instruments]. In: Manzi, J., Gonzalez, R. & Sun, Y. (eds.), La Evaluación Docente en Chile [The Chilean national teacher evaluation system], pp. 35-61. Pontificia Universidad Católica de Chile: Santiago, Chile.

Gaertner , H., & Pant, A. (2011, in press). How valid are school inspections? Problems and strategies for validating processes and results. Studies in Educational Evaluation.

Gates Foundation (2010). *The MET project framing paper.* Retrieved from http://www.gatesfoundation.org/highschools/Documents/met-framing-paper.pdf

Haertel, E. (2006). Reliability. In Brennan, R. (ed.), Educational Measurement, 4th ed., pp. 65-110. Westport, CT: Praeger Publishers.

Herman, J. & Baker, E. (2009). Assessment Policy: Making Sense of the Babel. In Sykes, G., Schneider, B., & Plank, D. (eds.), Handbook of Education Policy Research, pp. 176-190. New York: Routledge.

Herrera, J. (2009). Colegio de Profesores pide dejar sin efecto principales puntos de evaluación docente [Teacher Union asks to abandon key characteristics of the national teacher evaluation system]. La Tercera, September 22, 2009, p. 17.

Ingvarson, L., & Hattie, J. (Eds.). (2008). *Assessing teachers for professional certification: the first decade of the National Board for Professional Teaching Standards*. Oxford, UK: Elsevier.

Ingvarson, L., & Rowe, K. (2008). Conceptualising and evaluating teacher quality: Substantive and methodological issues. *Australian Journal of Education, 52*(1), 5–35. doi:10.1016/j.apmr.2010.02.005

Isoré, M. (2009). Teacher evaluation: Current practices in OECD countries and literature review. *OECD Education Working Papers*, *23*(1-49). doi:10.1787/223283631428

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

Kane, M. T. (2006). Validation. In Brennan, R. (ed.), Educational Measurement, pp. 17-64. Westport, CT: Praeger.

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

Kane, M., Crooks, T., & Cohen, A. (1999). Validating measures of performance. Educational Measurement: Issues and Practice, 18(2), 5-17.

Koretz, D. (2008). Measuring Up: What Educational Testing Really Tells Us. Cambridge, MA: Harvard University Press.

Koretz, D., Barron, S., Mitchell, K., & Stecher, B. (1996). Perceived effects of the Kentucky Instructional Results Information System (KIRIS). Santa Monica, CA: RAND.

Lane, S., & Stone, C.A. (2002). Strategies for examining the consequences of assessment and accountability programs. *Educational Measurement: Issues and Practice,* 21(1), 23-30.

Lane, S., & Stone, C. (2006). Performance Assessment. In Brennan, R. (ed.),

Educational Measurement, pp. 387-431. Westport, CT: Praeger.
Leal, P. & Santelices, V. (2010a). Análisis situación docentes eximidos y suspendidos sistema de Evaluación Docente [Analysis of the teacher´s status as exempt and suspended in the national teacher evaluation system]. Internal technical report MIDE UC, Pontificia Universidad Católica de Chile, Santiago, Chile.
Leal, P. & Santelices, V. (2010b). Análisis Decisiones tomadas por las Comisiones Comunales de Evaluación 2005-2008 [Analysis of decisions taken by the Local Evaluation Commissions 2005-2008). Internal technical report MIDE UC, Pontificia Universidad Católica de Chile, Santiago, Chile.
Linn, R. (2009). The concept of validity in the context of NCLB. In R. Lissitz (Ed.), *The concept of validity. Revisions, new directions and applications* (pp. 195 – 212). Charlotte, NC: Information Age Publishing.
Linn, R. (2006). Educational accountability systems (CSE Technical Report No. 687). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).
Linn, R. (1997). Evaluating the validity of assessments: The consequences of use. Educational Measurement: Issues and Practice, 16(2), 14-16.
Linn, R., Baker, E., & Dunbar, S. (1991). Complex, performance-based assessment: Expectations and validation criteria. Educational Researcher, 20(8), 15-21.
Manzi, J., González, R., & Sun, Y. (Eds.). (2011.). *La evaluación docente en Chile*. Santiago, Chile: Facultad de Ciencias Sociales, Escuela de Psicología, PUC.
Manzi, J., Strasser, K., San Martin, E., & Contreras, D. (2008, Febrero). Quality of education in Chile. Retrieved from: http://idbgroup.org/res/laresnetwork/files/pr300finaldraft.pdf
Manzi, J., Preiss, D., Gonzalez, R., Flotts, P. & Sun, Y. (2008). Design and Implementation of a National Project of Teaching Assessment: The Chilean Experience. Paper presented at the annual meeting of the American Educational Research Association, March 24-28, 2008, New York City, USA.
Messick, S. (1994). The Interplay of Evidence and Consequences in the Validation of Performance Assessments. *Educational Researcher, 23*(2), 13-23.
Messick, S. (1995). Standards of validity and the validity of standards in performance assessment. Educational Measurement: Issues and Practice, 14(4), 5-8.
Messick, S. (1989). Validity. In Linn, R. (ed.), Educational Measurement, 3rd ed., pp. 13-103. New York: MacMillan.
Milanowski, A. & Heneman III, H. (2001). Assessment of Teacher Reactions to a Standards-Based Teacher Evaluation System: A Pilot Study. Journal of Personnel Evaluation in Education, 15(3), 193-212.
Ministry of Education (2004). *Marco Para La Buena Enseñanza* [Guidelines for Good Teaching]. Santiago: Ministerio de Educación.
Ministry of Education (2003). *Attracting, Developing and Retaining Effective*

Teachers, OECD Activity, Country Background Report for Chile. Prepared by the Ministry of Education, Planning and Budget Division, Department of Research and Statistics. Retrieved on November 28, 2006, from: http://www.oecd.org/dataoecd/37/31/26742861.pdf

Mintrop, H., & Trujillo, T. (2005). Corrective action in low-performing schools: Lessons for NCLB implementation from first-generation accountability systems. Education Policy Analysis Archives, 13(48).

Moss, P. (2008). A critical review of the validity research agenda of the National Board for Professional Teaching Standards at the end of its first decade. In Ingvarson, L., & Hattie, J. (Eds.). (2008). *Assessing teachers for professional certification: the first decade of the National Board for Professional Teaching Standards*, pp. 257-312. Oxford, UK: Elsevier.

Myford, C. M., & Engelhard, G. (2001). Examining the psychometric quality of the National Board for Professional Teaching Standards Early Childhood/Generalist assessment system. *Journal of Personnel Evaluation in Education, 15*(4), 253–85.

National Research Council (2008). *Assessing accomplished teaching: Advanced - level certification programs*. Committee on Evaluation of Teacher Certification by the National Board for Professional Teaching Standards. Milton D. Hakel, Judith Anderson Koenig, & Stuart W. Elliott, editors. Washington, D.C.: The National Academies Press.

NCME Newsletter, vol. 18(1), March 2010.

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

Organisation for Economic Co-operation and Development (2004). Reviews of national policies for education: Chile. Organisation for Economic Co-operation and Development, Centre for Co-operation with Non-members: Paris, France.

Pecheone, R. L., & Chung, R. R. (2007). *PACT Technical report. Summary of validity and reliability studies for the 2003 - 04 pilot year*. Menlo Park, CA: Stanford University. Retrieved from Performance Assessment for California Teachers website: http://www.pacttpa.org/_files/Publications_and_Presentations/PACT_Technical_Report_March07.pdf

Pecheone, R. L., & Chung, R. R. (2006). Evidence in teacher education. The performance assessment for California teachers (PACT). *Journal of Teacher Education*, *57*(1), 22-36. doi:10.1177/0022487105284045.

Popham, W. (1997). Consequential validity: Right concern, wrong concept. Educational Measurement: Issues and Practice, 16(2), 9-13.

Preacher, K. & McCallum, R. (2003). Repairing Tom Swift's electric factor analysis machine. Understanding Statistics, 2 (1), 13-43.

Santelices, M. V., & Taut, S. (2011). Convergent validity evidence regarding the Chilean standards-based teacher evaluation system. *Assessment in Education: Principles, Policy & Practice*, *18*(1), 73-93. doi:10.1080/0969594X.2011.534948

Santelices, M. V., & Taut, S. (2010). Convergent validity evidence regarding the Chilean standards-based teacher evaluation system. Technical Report MIDE UC IT 1002, Pontificia Universidad Católica de Chile. Retrieved from: http://mideuc.cl/wp-content/uploads/2011/09/it1002.pdf

Santelices, V., Taut, S., Araya, C., & Manzi, J. (2009). Consequential Validity of Chile's Teacher Evaluation System: Consequences at the Municipal (Local) Level. *Paper presented at the Annual Meeting of the American Educational Research Association, April 13-19, 2009*, in San Diego, USA.

Santelices, M. V., Taut, S., & Valencia, E. (2009). Relación entre los resultados de la Evaluación Docente y los Planes de Superación Profesional: Estudio descriptivo. [Relationship between evaluation results and professional development plans: descriptive study] Internal document MIDE UC. Santiago, Chile: Pontificia Universidad Católica de Chile, Escuela de Psicología, Centro de Medición MIDE UC.

Santelices, V., Taut, S. & Valencia, E. (2008). Estudio exploratorio comparativo de los resultados de los programas AVDI, AEP y SEDD entre 2002 y 2006 [Exploratory comparative study of the results of AVDI, AEP and SEDD programs, between 2002 and 2006]. Retrieved from: http://www.mideuc.cl/docs/informes/it1007.pdf

Santiago, P. & Benavides, F. (2009). Teacher evaluation: A conceptual framework and examples of country practice. Paris, France: OECD. Retrieved from: http://www.oecd.org/dataoecd/16/24/44568106.pdf

Schafer, W., Wang, J. & Wang, V. (2009). Validity in action: State assessment validity evidence for compliance with NCLB. In R. Lissitz (Ed.), *The concept of validity. Revisions, new directions and applications* (pp. 173-193). Charlotte, NC: Information Age Publishing.

Shavelson, R. J., & Webb, N. M. (1991). Generalizability theory: A primer. Newbury Park, CA: Sage Publications.

Shepard, L. (1997). The centrality of test use and consequences for test validity. Educational Measurement: Issues and Practice, 16(2), 5-8, 13, 24.

Stecher, B., Barron, S., Chun, T., & Ross, K. (2000). The effects of the Washington state education reform on schools and classrooms (CSE technical report no. 525). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Strauss, A. L., & Corbin, J. M. (1990). *Basics of qualitative research: Grounded theory procedures and techniques*. Thousand Oaks, CA: Sage Publications, Inc.

Sun, Y., Correa, M., Zapata, Á., & Carrasco, D. (2011). Resultados: Qué dice la Evaluación Docente acerca de la enseñanza en Chile [Results: What does the Assessment Teaching say about teaching in Chile]. In J. Manzi, R. González, & Y. Sun (Eds.), *La evaluación docente en Chile* (pp 91-135). Santiago, Chile: Facultad de Ciencias Sociales, Escuela de Psicología, PUC.

Tabachnick, B. G., & Fidell, L. S. (1996). Principal components and factor analysis. In Using Multivariate Statistics, 3rd ed., pp. 635-707. New York: Harper-Collins.

Taut, S., Santelices, V., Araya, C. & Manzi, J. (in press). Perceived effects and uses of the national teacher evaluation system in Chilean elementary schools. Paper accepted for publication in a special issue of *Studies in Educational Evaluation*.

Taut, S., Santelices, V., Araya, C. & Manzi, J. (2011). Effects and uses of the national teacher evaluation system in Chilean elementary schools. *Paper presented at the Annual Conference of the American Educational Research Association, April 8-12, 2011*, New Orleans, USA.

Taut, S., Santelices, V., Araya, C. & Manzi, J. (2010). Theory underlying a national teacher evaluation program. Evaluation and Program Planning, 33, 477-489. Retrieved from: http://dx.doi.org/10.1016/j.evalprogplan.2010.01.002

Taut, S., Santelices, 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. Informe técnico MIDE UC IT1007. Disponible en http://www.mideuc.cl/docs/informes/it1007.pdf

Taut, S., & Santelices, V. (2007). Validating the Chilean National Teacher Evaluation System: A comprehensive research agenda. *Paper presented at the Annual Meeting of the American Educational Research Association, April 9-13, 2007*, in Chicago, USA.

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*(4), 132-142. doi:10.1016/j.stueduc.2011.02.002

Tornero, B. (2009). *Sistema de Evaluación del Desempeño Docente en Chile: Significados y Percepciones de Profesores Rebeldes.* Tesis para optar al grado de Magíster en Psicología Educacional, Pontificia Universidad Católica de Chile, Santiago, Chile.

Valencia, E. & Taut, S. (2008). Estudio de dimensionalidad del portafolio de Docentemás 2007 [Dimensionality study of the Docentemás portfolio 2007]. Informe técnico MIDE UC, Pontificia Universidad Católica de Chile..

Wolming, S., & Wikström, C. (2010). The concept of validity in theory and practice. Assessment in Education: Principles, Policy and Practice, 17(2), 117-132.

Appendices

Table 1.
Validity matrix of evidence for NTES.

Types of evidence related to	Available studies or evidence	Pending studies
1. Content	Correspondence table MBE-DM instruments, all years (**2005 correspondence MBE- DM portfolio scoring rubric)**	External, independent study on content validity of all assessment instruments
	Expert teacher commissions design portfolios (instructions and rubrics) based on MBE, all years	External, independent study on MBE quality (although MBE is almost literal translation of Danielson framework)
2. Response process	Think-aloud pilot studies of portfolio and peer interview instruments, all years	No such studies for IRT and self-evaluation
3. Internal structure	**Exploratory factor analyses, all years**	Confirmatory factor analyses
4. Relations to other variables		
Convergent	**Correspondence with AEP program, 2004-2006**	Studying teaching practice of basic versus competent teachers
	In-depth study of teaching practice of outstanding versus unsatisfactory teachers, 2006	Studies using longitudinal achievement data of students whose teachers have been evaluated by DM
Criterion	Correlational studies NTES-SIMCE	HLM analyses using student gain scores matched with teachers' NTES results
5. Consequences		
Addressing diagnosed weaknesses	**PSP studies looking at PSP quality and impact (using DM reevaluation score as criterion)**	Studies on quality of reports
	Evidence from focus groups and interviews with teachers and at local level	Impact on student achievement
Providing incentives to increase job satisfaction	**AVDI studies**	
	Recognition practices at school and local level	
	Interviews with teachers	
Fostering peer collaboration	**Evidence from interviews with school leaders and teachers**	
Informing educational decision-making	**Evidence from interviews and surveys of local authorities and school leaders**	
Improving student achievement	Pending (sufficient longitudinal student data not available so far)	
Unanticipated negative effects	**Evidence from interviews at all levels**	
Unanticipated positive effects	**Evidence from interviews at all levels**	

Table 2.
Matrix of evidence of other aspects of technical quality related to NTES' validity.

Other aspects of technical quality influencing validity	Available studies or evidence	Pending studies
Scaling, equating & standard setting	Based on professional judgment of expert teacher commissions, measurement team and empirical results of respective evaluation process, all years	Review of these processes by external measurement experts
Assessment construction, administration & scoring	Descriptions of careful construction and scoring processes, all years	Review of these processes by external measurement experts
Reliability & generalizability	Reliability coefficients (internal consistency) for each evaluation instrument, all years	IRT-based item analysis
	Generalizability studies on rater effect, all years, and occasion, 2010	G-studies including other facets
Fairness	**Functioning of local evaluation commissions, 2009**	DIF studies
	Local exemption and suspension practices, 2009	Systematic information on cheating and other unethical conduct
	Mean differences between sub-groups of evaluatees to detect bias, all years	
	Study of item bias, preliminary (2010)	

Table 3. Loadings from Exploratory Factor Analysis of 2009 NTES Portfolio (factor loadings equal or larger than 0.2).

Rotated Factor Matrix

Items	Factor					
	1	2	3	4	5	6
a1					.478	
a2	.231				.708	
a3	.230				.348	.221
b1	.429					
b2	.600					
b3	.632					
c1			.883			
c2			.868			
c3	.205		.493		.284	
d1	.400					
d2	.484					
e1	.458					
e2	.554					
e3	.521					
f1		.247		.645		
f2		.551				
f3		.263		.721		
g1		.451		.239		
g2		.473				
g3		.311		.401		
h1		.480				.353
h2		.580				
h3		.631				.339
h4		.355		.212		.565

Note: Maximum likelihood extraction method with varimax rotation

Table 4. Loadings from Confirmatory Factor Analysis for 2010 NTES Portfolio.

	Item	Estimate	S.E.	Est./S.E.	Two-Tailed P-Value
DimA	A1	0.307	0.014	22.453	0.000
	A2	0.659	0.013	50.174	0.000
	A3	0.466	0.013	36.506	0.000
DimB	B1	0.553	0.012	47.614	0.000
	B2	0.516	0.012	43.363	0.000
	B3	0.489	0.012	42.376	0.000
DimC	C1	0.937	0.005	177.098	0.000
	C2	0.894	0.005	166.433	0.000
	C3	0.595	0.008	76.004	0.000
DimD	D1	0.484	0.013	36.114	0.000
	D2	0.547	0.014	39.583	0.000
DimE	E1	0.615	0.01	60.522	0.000
	E2	0.534	0.011	49.716	0.000
	E3	0.596	0.01	59.102	0.000
DimF	F1	0.5	0.016	31.379	0.000
	F2	0.559	0.014	39.573	0.000
	F3	0.603	0.018	34.289	0.000
DimG	G1	0.526	0.011	46.461	0.000
	G2	0.57	0.011	52.35	0.000
	G3	0.494	0.012	42.274	0.000
DimH	H1	0.616	0.01	60.327	0.000
	H2	0.567	0.011	51.411	0.000
	H3	0.638	0.012	53.147	0.000
	H4	0.573	0.011	52.24	0.000

Note: WLSMV estimation method; RMSEA=0.046; CFI=0.955; TLI=0.963

Table 5.

Internal consistency of NTES instruments 2005-2010

Instrument	2005	2006	2007	2008	2009	2010
Portfolio, written part	0.81	0.80	0.74	0.78	0.76	0.77
Portfolio, video-taped lesson	0.79	0.75	0.70	0.74	0.72	0.72
Supervisor assessment	0.97	0.99	0.98	0.96	0.98	0.97
Peer interview	0.80	0.87	0.82	0.81	0.80	0.78

Table 6.
Effect sizes (Cohen's d) for t-tests that showed significant mean differences.

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

Table 7.
Teacher participation and results in AVDI test

Year	Eligible teachers	Non-participating teachers	Participating teachers	Participation rate	Unsuccessful applicants	Successful applicants	Proportion of AVDI attainment*	Proportion of competent and outstanding in the AVDI test*	Applicants receiving 25% salary bonus
2004	2425	1234	1191	49%	871	320	27%	27%	9
2005	3182	2084	1098	35%	828	270	25%	25%	9
2006	6329	3089	3240	51%	865	2375	73%	23%	14
Total	11936	6407	5529	46%	2564	2965	54%	23%	32

Note. *Over participating teachers

Table 8.

Teacher status in 2009 of teachers evaluated 2003-2008.

Status (2009)	Unsatisfactory at least once	Competent at least once	Outstanding at least once	Competent or outstanding at least once
Teaching in municipal schools	68.1%	88.3%	92.1%	88.8%
Teaching in private-subsidized schools	3.6%	1.9%	1.5%	1.9%
Inactive	31.9%	11.7%	7.9%	11.2%
Total	100%	100%	100%	100%