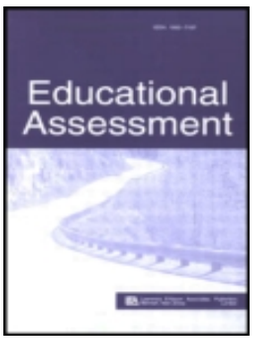

Routledge
Taylor & Francis Group

Educational Assessment

ISSN: 1062-7197 (Print) 1532-6977 (Online) Journal homepage: http://www.tandfonline.com/loi/heda20

Validation of a National Teacher Assessment and Improvement System

Sandy Taut , María Verónica Santelices & Brian Stecher

To cite this article: Sandy Taut , María Verónica Santelices & Brian Stecher (2012) Validation of a National Teacher Assessment and Improvement System, Educational Assessment, 17:4, 163-199, DOI: 10.1080/10627197.2012.735913

To link to this article: https://doi.org/10.1080/10627197.2012.735913

Educational Assessment, 17:163–199, 2012

ISSN: 1062-7197 print/1532-6977 online
DOI: 10.1080/10627197.2012.735913

Routledge
Taylor & Francis Group

Validation of a National Teacher Assessment and Improvement System

Sandy Taut and María Verónica Santelices
Pontificia Universidad Católica de Chile

Brian Stecher
RAND

The task of validating a teacher assessment and improvement system is similar whether the system operates in the United States or in another country. 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 6 years the authors have performed validation research on NTES using a variety of methods and data sources. This article describes our validation research agenda, the results of major validation studies, and an integration of the existing evidence, and it offers the authors' preliminary judgment about NTES's validity. The article also offers a critical reflection regarding the decisions taken while driving the long and winding validation road, and the lessons we learned during this politically and methodologically complex journey.

Improving teacher effectiveness has become a major focus of educational reform policy in the United States. However, existing efforts to measure teacher effectiveness and use the information as a tool to improve student performance have only recently started a more thorough evaluation. The experiences of educators in Chile, who are implementing and validating a national teacher evaluation system (NTES), can provide useful examples for researchers and policymakers in the United States.

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

Correspondence should be sent to Sandy Taut, Escuela de Psicología, Centro de Medición MIDE UC, Pontificia Universidad Católica de Chile, Av. Vicuña Mackenna 4860, Macul, Santiago, Chile. E-mail: staut@ucla.edu

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). The federal government supports and encourages these policies through the 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). 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.

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 "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 [AERA, APA & NCME], 1999; Kane, 2006; Linn, 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, a 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 validity of such a system is an equally complex process (Gaertner & Pant, 2011; Koch & DeLuca, 2012, Wolming & Wikström, 2010).

The scope of the effort may explain why comprehensive validation efforts regarding large-scale teacher 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 (Ingvarson & Hattie, 2008; National Research Council, 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 (Linn, 2009; Schafer, Wang, & Wang, 2009).

This article 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 NTES. The first two authors were based in the research department of the university measurement center responsible for developing and implementing NTES. For the past 6 years we have been performing validation research on NTES using a variety of methods and data sources.

The article 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 6 years. The heart of the article reviews the evidence assembled to address the various validity questions, ranging from content through consequences. We close with a preliminary judgment about NTES's 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

Introduction of the NTES

The Chilean NTES was introduced by the Ministry of Education amid some controversy in 2003, and since 2005 it has been mandatory for teachers in municipal schools nationwide. Prior to the NTES, the teacher evaluation system in Chile was locally implemented, was related to seniority more than to performance, was subjective in nature, and had little to do with a teacher's classroom performance; the results often had few 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 years of 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

Broadly speaking, the NTES was designed to inform educational policy at the national level by providing overall information on the quality of teachers in the country's public schools. It was also designed to improve the overall 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 hold teachers accountable through personnel actions. Thus, the NTES evaluation system serves both summative and formative purposes, and it operates at multiple levels—national, municipal, school, and individual.

On the summative side, those teachers receiving an "unsatisfactory" result have to reevaluate the following year and if they repeat their "unsatisfactory" performance two years in a row they are subject to loss of employment. Likewise, "basic" teachers have to reevaluate after two years and if they get three "basic" performance ratings in a row they are also subject to removal.[1] 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 4 years until their reevaluation (see Figure 1 for a summary of NTES's legally prescribed consequences). On the formative side, individual reports are sent to each evaluated teacher, indicating strengths and weaknesses on a number of dimensions, defining the expected performance for each dimension and contrasting this information with the observed performance. Low-performing teachers are offered professional development in their respective municipalities to address the weaknesses that the NTES has diagnosed for each teacher. The

[1]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 4 years, instead of after 2 years, and there were no punitive consequences attached to repeated basic performance.

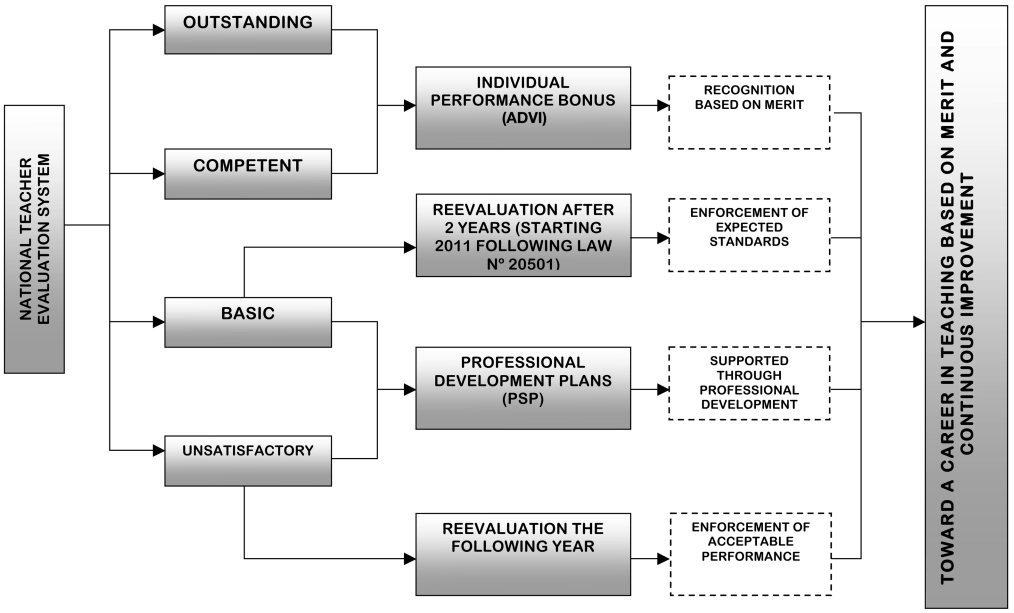

FIGURE 1 National Teacher Evaluation System's consequences according to new 2011 legislature.

professional development activities are monitored via an online tool used by municipalities, implementers, and teachers. The NTES also 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 as well.

The first step in validating a testing program, like NTES, is to clarify how the test scores will be interpreted and the purposes for which they will be used. In formal terms this is referred to as developing the interpretive argument for the use of the assessment (Kane, 2006). As we noted, the creation of NTES was contentious, and as a result, the interpretive argument for the NTES is complex. 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 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 purposes for NTES results; 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, and improving teachers' job prospects by offering monetary incentives (for details, see Taut, Santelices, Araya, & Manzi, 2010). In short, the stakeholders recognized that the NTES results would serve two broad purposes: providing formative data on individual teachers to improve their practice, and providing summative data to support individual teacher rewards and sanctions. To support these goals the test scores will be interpreted both as an indicator of a teacher's overall competence and as a guide to identifying areas for improvement. To be interpreted in this manner, NTES scores must correctly distinguish overall teacher effectiveness, particularly at the top and bottom levels, and they must reasonably identify teachers strengths and weaknesses.

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* (MBE)" [Guidelines for Good Teaching] (Ministry of Education, 2004). These professional teaching standards were developed on the basis of Danielson's (1996) Framework for Teaching. 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." The MBE formally establishes the dimensions of good instruction and serves as the basis for developing the NTES instruments and for interpreting the NTES scores.

Components of the NTES

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

Portfolio assessment. The portfolio asks teachers to describe lesson planning and classroom evaluation materials for a specific, predefined 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 min) from each teacher is videotaped by an external contractor. Portfolio instructions and specific items of the scoring rubrics differ across years, whereas the general structure of the portfolio and the eight dimensions of the scoring rubric remain constant across time. The eight dimensions are as follows: 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 videotaped lesson. Each of the eight dimensions of the portfolio rubric is operationalized by three items, so the rubric contains a total of 24 items.

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. Since 2010, 2-hr trainings are provided to a subsample of municipal evaluation coordinators and school leaders. In 2010, about 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 2-day training (about 1,400 are preselected, of which the 10% worst performing are asked to leave the process). This process is intended to ensure valid and reliable evaluation data and 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 instrument is responded to or scored by raters on a 4-point scale: 1 (*unsatisfactory*), 2 (*basic*), 3 (*competent*), and 4 (*outstanding*). Performance-level descriptions are developed sequentially every year for each item starting by defining "competent" performance, followed by defining "basic," then "unsatisfactory," and last "outstanding" performance. Operationalized definitions of "competent" performance are extrapolations of the teaching standards and based on the professional judgment of the pedagogical and disciplinary experts involved in NTES.

In the case of the supervisor and self-assessments, the items are responded to directly on the 4-point scale, whereas 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 4 points on the scale for each item. The notes taken during the peer interview are later scored by the same interviewer, whereas 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.

Because portfolio scoring is the most complex part of the evaluation and contributes most weight to the overall score (see next), we focus attention on this aspect. Each year the portfolios and corresponding scoring rubrics are developed based on the following sources of information:

1. The official teaching standards (Marco para la Buena Enseñanza, MBE).
2. Videos of teaching practice in each subject and grade level are consulted to make sure the portfolio links to actual classroom practice.
3. 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.

4. Pilot studies are implemented for each year's draft portfolios, applying the think-aloud technique both individually and in groups of teachers.
5. Pilot studies are implemented for the final version of the portfolios resulting in about 12 completed portfolios for each subject and grade level, to be used for constructing the scoring rubrics and for rater training.
6. 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 hr, to get to know the scoring rubric and learn to apply it by using the same practice portfolios, whereas their performance is being monitored constantly. The professionals in charge of supervising the rating process also receive training (40 hr) during the month prior to the scoring period. There is also a 3-day trial period when all processes are tested and raters continue practicing and being monitored. Every Monday during the 3-week scoring period, all raters complete a group scoring session with their supervisors. Also, 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.

According to NTES data of 66,938 teachers obtaining final evaluation results between 2003 and 2010, the majority (60%) of evaluated teachers receive the performance categorization of "competent," whereas about one third are evaluated as showing "basic" performance (30%). Only a small percentage (8%) is evaluated as "outstanding," and even fewer (2%) 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 4 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 1st year of their teaching career, work as peer evaluators, or since 2007 are within 3 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, 2010a).

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 is charged with confirming or modifying the final NTES result of each teacher evaluated in the respective municipality, taking into account local contextual information provided by the evaluated teacher and his or her supervisors. Each modification must be justified in an online database. So far only about 3% of final scores are modified every year (Manzi, González, & Sun, 2011; Leal & Santelices, 2010b).

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 or 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 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 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*, or 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 4 years until their reevaluation. 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 from 5% to 15% to 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*, or PSP) 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 online tool.

The MBE standards form the basis not only for the NTES but also for a voluntary accreditation of teaching excellence program (*Asignación de Excelencia Pedagógica*, or AEP), which was introduced in 2003. Although these two programs assess similar indicators of competent teaching, the instrument development and scoring processes are independent of each other. AEP engages in a standard-setting process by subject matter and separate for elementary, middle, and high school teaching. The AEP assesses teachers on the basis of a portfolio including 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 just mentioned. 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 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 Standards (AERA, APA & NCME, 1999) 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" (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" (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).

Although 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 (National Council on Measurement in Education, 2010). Kane (2006) 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), Linn (2009), Messick (1989, 1995), and Shepard (1997)

on including an evaluation of both the meaning of test scores and the consequences of their use within our research program on NTES's validity.

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 (Bond, Smith, Baker, & Hattie, 2000; Moss, 2008; National Research Council, 2008). 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.

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 noted that "certification programs generally rely on content-based validity evidence," whereas 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.

The National Research Council (2008) further explained that although external criteria for validating certification tests are difficult to find and even unnecessary, the NBPTS has a predictive component related to student achievement, and they dedicated 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).

DECISIONS WE TOOK IN DEFINING OUR VALIDATION RESEARCH AGENDA

In 2005 we started investigating the validity of the NTES. As previously noted, we sought to delineate the interpretive argument that needed to be made for the NTES scores, and then we developed a long-term research agenda to gather evidence to establish whether those claims were reasonable, that is, we began to develop the validity argument (Kane, 2006). We used the Standards (1999) as a starting point to structure our data collection and analysis to support the formative and summative interpretations of NTES scores we identified, and we organized our work around the types of evidence delineated in the standards, that is, 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 main ideas.

In Tables 1 and 2 we present summaries of all available and still pending evidence for the NTES. Although Table 1 focuses on validity evidence, Table 2 reflects other relevant technical aspects of the NTES assessment system: reliability and fairness. As previously mentioned, in this article we report primarily the studies with which we were directly involved, and those studies are shown in bold in Tables 1 and 2.

TABLE 1
Validity Matrix of Evidence for NTES

Types of Evidence Related to	*Available Studies or Evidence*	*Pending Studies*
1. Content	Correspondence table MBE-NTES instruments, all years (**2005 correspondence MBE-NTES portfolio scoring rubric**)	a (p) External, independent study on content validity of all assessment instruments b (p) 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	2 (p) Think aloud pilot studies for supervisor assessment and self-evaluation
3. Internal structure	**Exploratory and confirmatory factor analyses of portfolio**	3 (p) Confirmatory factor analyses of supervisor assessment, peer interview and self-evaluation
4. Relations to other variables		
Convergent	a. **Correspondence between NTES and AEP programs, 2004–2006**	
	b. **In-depth study of teaching practice of outstanding versus unsatisfactory teachers comparing NTES performance to performance on alternative measures, 2006**	a (p) Studying teaching practice of basic versus competent teachers
Criterion	c. **HLM analyses using student gain scores matched with teachers' NTES results, 2006**	b (p) Studies using longitudinal achievement data of students whose teachers have been evaluated by NTES
	d. Correlational studies NTES-SIMCE	
5. Consequences		
Addressing diagnosed weaknesses	a. **Studies looking at PSP implementation quality and re-evaluation scores of PSP participants**	a (p) Studies on quality of reports
	b. **Evidence from focus groups and interviews with teachers and at local level about PSP quality and impact**	b (p) Impact of PSP program on teaching practice in the classroom and student achievement
Providing incentives to increase job satisfaction	c. **AVDI participation and success rate studies**	
	d. **Reports on recognition practices at school and local levels based on interviews and focus groups**	
	e. **Focus groups with teachers about incentives policies**	
Fostering peer collaboration	f. **Interviews with school leaders and teacher focus groups about changes in peer collaboration**	
Informing educational decision making	g. **Interviews and surveys of local authorities and school leaders about their use of NTES results**	
	h. **Study of job trajectories of unsatisfactory teachers**	
Unanticipated negative effects	i. **Evidence from interviews at all levels**	
Unanticipated positive effects	j. **Evidence from interviews at all levels**	

Note. Bold indicates the studies with which we were directly involved. Pending studies are designated by (p). MBE-NTES = *Marco para la Buena Enseñanza*–National Teacher Evaluation System; AEP = *Asignación de Excelencia Pedagógica*; HLM = Hierarchical Linear Modeling; SIMCE = a curriculum-based standardized test for Chilean students; PSP = *Planes de Superación Profesional*; AVDI = *Asignación Variable por Desempeño Individual.*

TABLE 2
Matrix of Evidence of Other Aspects of Technical Quality Related to the National Teacher Evaluation System's Validity

Other Aspects of Technical Quality Influencing Validity	*Available Studies or Evidence*	*Pending Studies*
6. 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
7. Assessment construction, administration, & scoring	Descriptions of careful construction and scoring processes, all years	Review of these processes by external measurement experts
8. Reliability & generalizability	a. Reliability coefficients (internal consistency) for each evaluation instrument, all years	IRT-based item analysis
	b. **Generalizability studies on rater effect, all years, and rating occasion, 2010**	G-studies including other facets
9. Fairness	a. **Functioning of local evaluation commissions, 2009**	DIF studies
	b. **Local exemption and suspension practices, 2009**	Systematic information on cheating and other unethical conduct
	c. Mean differences between sub-groups of evaluatees to detect bias, all years	
	d. **Study of item bias, preliminary, 2010**	

Note. Bold indicates the studies with which we were directly involved. IRT = item response theory; DIF = differential item functioning.

Limitations

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's validity, reliability and fairness (Anastasi, 1986; Cronbach, 1989; 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 previously explained, we focus on the most important purposes and uses (Herman & Baker, 2009). Second, we decided to focus on (a) the overall assessment score because it has direct consequences for individual teachers, schools, and municipalities, and (b) 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 et al., 2011).

Research Questions

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

that provided the most important evidence regarding the interpretation of NTES' overall scores as indicators of teacher performance and subscale scores as guides to improvement.

- 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 municipalities?
- 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 apply 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 (Leal & Santelices, 2010; Santelices, Taut, Araya, & Manzi, 2009; Santelices, Taut, & Valencia, 2008, 2009; Taut, Santelices, Araya, & Manzi, 2011a; Taut, Santelices, & Valencia, 2010; Valencia & Taut, 2008) and related journal articles (Santelices & Taut, 2011; Taut, Santelices, Araya, & Manzi, 2010, 2011b; Tornero & Taut, 2010). 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 coauthors 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). Whereas 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 preestablished underlying portfolio structure, which in our case corresponds to the eight dimensions as eight related but distinguishable underlying constructs (Preacher & McCallum, 2003; Tabachnick & Fidell, 1996). Our exploratory factor analyses applied the Principal Axis Factoring and Maximum Likelihood estimation methods, as well as various factor retention rules (rule of Kaiser-Gutmann, screen test, interpretability). We used rotation method Oblimin, however, reaching similar findings when applying Varimax (orthogonal) rotation. The confirmatory factor analysis was based on 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. To evaluate model fit we consulted various indices: chi-square test of model fit, comparative fit index (CFI), Tucker-Lewis fit index (TLI), and root mean square error of approximation (RMSEA). According to Brown (2006) the appropriate cutoff values are >0.95 for CFI and TLI, and <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. To study teacher performance with other measures, 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; Santelices et al., 2008; Taut & Santelices, 2007). Because of funding restrictions, this study was conducted with a sample of 58 highest and lowest scoring NTES teachers, as these were the assessment categories that faced the most consequences at the time, namely, the loss of employment and the opportunity to get a substantial salary bonus. Similar analyses considering teachers evaluated as "basic" and "competent," the two adjacent and more numerous categories, are currently under way.

We studied the consequences of the teacher assessment system by applying a mix of methods: First 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 (Santelices, Taut, Araya, et al., 2009; Santelices, Taut & Valencia, 2008, 2009; Taut, Santelices, Araya, & Manzi, 2010, 2011b; Taut, Santelices, & Valencia, 2010; Tornero & Taut, 2010).

During the interviews and group discussions we set out asking general, nonleading questions regarding the programs' consequences first, and only as a second step did we probe more specific intended effects and uses. We interviewed 19 municipal actors in 10 purposively sampled municipalities, conducted interviews with 57 school leaders from 30 public elementary schools from those same 10 municipalities, implemented nine focus group discussions with a total of 46 teachers, and interviewed 10 teachers who received an "unsatisfactory" performance rating as well as nine teachers who had at least once actively refused to participate in NTES. We transcribed all interviews and conducted content analysis using ATLAS.ti software. We

developed a codebook based on initial open coding, consolidating the code list later by forming code families, refining code definitions, and adding new codes if necessary. All interviews were double-coded, resulting in one consented version of each coded interview. At the municipal level we studied salience as well as the strength, existence, or absence of the intended and unintended consequences by municipality. In analyzing the data at the school level we distinguished between a code's frequency, presence/absence, and salience. At the teacher level we studied frequency, presence, and absence of each code considering the specificity, emotionality, and extension of the discourse. The information from interviews was then summarized in municipality-level, school-level and teacher-level reports.

We also calculated indices reflecting the interrater reliability and generalizability of the portfolio scoring process (Haertel, 2006). Interrater reliability is calculated as an intraclass correlation, ranging from zero to 1.00. Computing an analysis of variance provides the Mean Square for persons (MS_p) and for error (MS_e).

$$R_i = \frac{MS_p - MS_e}{MS_p + MS_e(K-1)} \tag{1}$$

where i indicates that the average reliability of a single rater is estimated and K corresponds to the number of judges rating each object (Shavelson & Webb, 1991). 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 rating occasion as a facet.

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

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 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 are assessed by other parts of the NTES, the peer interview, the supervisor assessment, and the self-assessment.[2] In Domain A, in particular, there is limited

[2] 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.

coverage of standards related to subject-specific pedagogy[3] 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, as 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[4]

As mentioned previously, the scoring rubric for the NTES portfolio contains 24 indicators grouped in eight dimensions. Since 2005 we routinely performed exploratory factor analysis of every year's NTES portfolio data and in 2010 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 videotaped lesson change from year to year, but in some years they have neatly re-created 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, 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 five dimensions of the written part of the portfolio.

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

[4]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 videotaped lesson, between 0.96 and 0.99 for the supervisor assessment, and between 0.78 and 0.87 for the peer interview.

TABLE 3
Loadings from Exploratory Factor Analysis of 2009 National Teacher Evaluation System Portfolio

	Rotated Factor Matrix					
	Factor					
Items	*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. Factor loadings are equal or larger than 0.2.

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

TABLE 4
Loadings from Confirmatory Factor Analysis for 2010 National Teacher Evaluation System Portfolio

	Item	*Estimate*	*SE*	*Est./SE*	*Two-Tailed p Value*
Dim A	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
Dim B	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
Dim C	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
Dim D	D1	0.484	0.013	36.114	0.000
	D2	0.547	0.014	39.583	0.000
Dim E	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
Dim F	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
Dim G	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
Dim H	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. Weighted least squares means and variance adjusted estimation method; root mean square error of approximation = 0.046, comparative fit index = 0.955, Tucker–Lewis index = 0.963.

teacher performance with student achievement as measured by standardized tests (SIMCE).[5] The third study explored the relationship between teachers' NTES results and their scores in the Pedagogical Excellence Certification program (AEP), which also involves a portfolio based on the Guidelines for Good Teaching (MBE) (Santelices et al., 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:

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

- 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-min 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 2-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" teachers on half of the performance indicators and showed positive but not significant differences on the remaining indicators (see Table 5 for details on the performance indicators and the effect sizes). We found especially strong and practically significant differences related to time on task during lessons, lesson structure, student behavior, and classroom assessment materials. We also found moderate correlations (r ranging from 0.3 to 0.6, $p < 0.05$) between the results these same teachers had obtained 1 year earlier in NTES and some indicators from the classroom observations, teaching materials binder and teachers' standardized test performance. These indicators were most strongly correlated to the NTES portfolio results. Classroom observations and binder assessments also correlated significantly with the NTES supervisor and peer evaluations. The NTES self-evaluation results showed the lowest correlations with

TABLE 5
Effect Sizes (Cohen's *d*) for t Tests That Showed Significant Mean Differences

Instrument	*Indicator/Subscale*	*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
Postobservation 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

the performance on the validity study's instruments. We interpret these results as evidence for the validity of the NTES to differentiate among extreme groups of teachers based on their pedagogical practices in the classroom. It is important to note, however, that our results are limited by the small sample of teachers studied and by the fact that the study took place the year following the NTES evaluation. This is problematic, especially for the lowest performing group of teachers, as NTES is supposed to motivate improvement by mandating participation in professional development and reevaluation the following year. However, in terms of differences among the highest and lowest performing teachers, such training and additional motivation would only have made it more difficult for our study to detect differences between the groups.

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 ($N = 1{,}044$ students of 40 teachers). These data were analyzed using hierarchical linear modeling and showed that teacher performance in NTES is a significant predictor of student achievement at the end of the school year ($p < .01$), while controlling additionally only for student achievement at the beginning of the year (Santelices & Taut, 2011). The limited sample size did not allow us to include students' or schools' socioeconomic variables or control for grade level or discipline. Although the level-2 variance component decreases by 14.34 as a result of the introduction of NTES performance in the model, the difference in explained variance between the unconditional and the conditional models is not substantial in terms of size. The overall deviance of the conditional models (8734.65) suggests that a significant proportion of the overall variance is not explained by the variables included in the model.

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, 2008; Ministry of Education, 2008, 2009). Their results tend to support the positive relationship between teacher performance in NTES and SIMCE achievement of the students with whom these teachers worked. 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 et al., 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 ($r = 0.33$), as well as between NTES final score and AEP portfolio score ($r = 0.36$). 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 (1 or 2 years apart) and by differences in the nature of participating in them: Although the NTES is mandatory and offers economic incentives that need an additional application and assessment process, 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's validity when considering the AEP assessment as criterion.

Evidence Related to the Consequences of the Teacher Assessment

We examined empirically whether NTES's intended consequences, as identified by our study of legal documents and stakeholder perceptions (see preceding details), were in fact observed, and what were its unintended consequences.

Our empirical findings indicate that the assessment program achieves some of its intended consequences while falling short on others (Santelices, Taut, Araya, et al., 2009; Santelices, Taut, & Valencia, 2008, 2009; Taut, Santelices, Araya, & Manzi, 2011a,b; 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 subgroups of interviewees was the experience 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 revise their teaching practice.

On the other hand, professional development was not effective according to many teachers (Cortés, Taut, Santelices, & Lagos, 2011; Santelices, Taut, & Valencia, 2009). 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 the district, which is to be spent on these teachers' professional development, as organized by the municipality. Professional development at the municipal level is planned 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 diverse 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 3-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 reevaluation had average scores above those of unsatisfactory teachers who did not take part in PSP ($N = 12$). 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). 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, two 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 subeffect 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 one unsatisfactory performance rating are three times as likely to leave the public teaching force than those who have never received such a result (32% vs. 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 second or 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 of 10 municipalities.

TABLE 6
Teacher Employment Status in 2009 of Teachers Evaluated 2003 to 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%

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 et al., 2008; see Table 7). We analyzed data from 2004 to 2006 regarding the participation of teachers in the incentives program associated with NTES. To qualify for a salary bonus, teachers have to pass a subject knowledge test corresponding to the subject matters they teach. If they show "outstanding" performance in both the NTES and the AVDI test, teachers 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 reevaluation 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 3 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 3 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 was 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 that psychological and motivational support 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, 2010). 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, 1990). Our findings indicated that several factors influenced the active rejection of the NTES by this group of teachers, including strong negative emotions

TABLE 7
Teacher Participation and Results in AVDI Test

Year	*Eligible Teachers*	*Nonparticipating Teachers*	*Participating Teachers*	*Participation Rate*	*Unsuccessful Applicants*	*Successful Applicants*	*Proportion of AVDI Attainment* [a]	*Proportion of Competent and Outstanding in the AVDI Test* [a]	*Applicants Receiving 25% Salary Bonus*
2004	2,425	1,234	1,191	49%	871	320	27%	27%	9
2005	3,182	2,084	1,098	35%	828	270	25%	25%	9
2006	6,329	3,089	3,240	51%	865	2,375	73%	23%	14
Total	11,936	6,407	5,529	46%	2,564	2,965	54%	23%	32

Note. AVDI = *Asignación Variable por Desempeño Individual.*
[a]Over participating teachers.

generated by NTES (such as fear of getting a poor result), cultural aspects of the teaching profession in Chile, and 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 a 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 interrater 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 ($R = 0.8$). Although 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 as "teachers"), (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).

The G-studies have found this "residual" error term 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, grade level and subject matter analyzed.

In 2010 we designed and implemented a G-study independent of the actual portfolio scoring process in which we included rater and scoring occasion (two times, 7 days apart) as facets in a completely crossed design; also, we compared the generalizability of the rater facet in two different scoring schemes: (a) Each rater scored the complete written module of a single teacher (scoring by teacher) and (b) each rater scored Dimension 1 in all of the portfolios assigned to him or her, then he or she continued with Dimension 2, and so on (scoring by dimension). The results indicated that 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).

In addition, 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 is some evidence 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 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 and eliminate consistently misfitting raters) we think more could be done. Possible modifications include progressing toward 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's 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 1 year of experience, teachers who are serving as a peer evaluator that same year, and teachers within 2 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 evaluated in their respective community. 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 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 (between 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 article 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, that is, we gathered multiple types of evidence related to validity, including consequences (AERA, APA, & NCME, 1999). Inspired by the guidelines of mixed methods research (Creswell & Plano Clark, 2007), in this section we attempt to summarize the evidence regarding the validity research questions we set out to answer and integrate all types of evidence to come to a final conclusion about NTES's validity.

Does the Teacher Evaluation Cover the Contents of Its Underlying Standards, the Guidelines for Good Teaching (MBE)?

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.

Do the Portfolio Scores Represent the Different Aspects (Factors) of Teaching Performance That NTES Reports Communicate for Formative Purposes?

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.

Does the Evidence About Relations With Other Variables Support or Weaken the Validity of the NTES?

We believe that the relationship between NTES results and other variables support the validity of the NTES final category. The first research question addressed the differences in teaching practice between the highest and lowest performing teachers as measured by alternative instruments using both similar and complementary indicators. The NTES process reveals real differences between these two groups of 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 and involving data from only 1 year, the study showed that teachers' evaluation result was a significant predictor of students' posttest score while controlling for students' pretest 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.

Does NTES Have the Intended Consequences, and Are There Important Negative, Unintended Consequences for Teachers, Schools, and Municipalities?

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. 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 professional development program, PSP), and improving teachers' work prospects by granting economic incentives associated with good individual performance (via incentives program,

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. A new Law (No. 20.501), in effect since 2011, does increase the bonuses substantially. 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 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 nonpaid 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 American Education Research Association, 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 because 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.

Is Portfolio Scoring Free From Rater Bias—or Is There a Substantial Amount of Variance in Assessment Scores Due to Rater Effect?

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 100%. Our analyses indicated that generalizability indexes could be improved substantially in this way. In addition, and because 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. Future research should also address other potential sources of error, such as assessment occasion.

Is the Evaluation Process Fair to All Evaluated Teachers, With Respect to the Role of the Local Evaluation Commissions in Modifying the Final Score, and the Right to Suspend the Evaluation?

"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.

Preliminary Conclusion About NTES's Validity Based on Evidence Amassed to Date

Given all our work on validating NTES so far, Koretz's (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's validity are judged is a matter of perspectives and values (Herman & Baker, 2009). Although 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's 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's 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; Correnti & Martinez, 2012; 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.

Integrating the Standards-Based Framework Organized by Types of Validity Evidence With an Argument Approach to Validation

Before discussing pending issues, lessons and challenges regarding NTES's validation, we present our attempt to translate our validation approach based on types of validity evidence (AERA, APA, & NCME, 1999) into an argument approach to validation (Kane, 2006). The two main purposes for the NTES are (a) to distinguish among teachers based on effectiveness and (b) to use this information to support teacher improvement. Table 8 summarizes the elements of an interpretive argument in support of these two purposes using a framework suggested by Kane, Crooks, and Cohen (1999) and shows how the research evidence previously presented relates to this structured argument (also see Bell et al., 2012; Hill, Kapitula, & Umland, 2011). We hope that looking at our work from this new angle contributes additional robustness to the conclusions we have presented here and helps us identify critical evidentiary gaps to be filled in the future.

TABLE 8
Evidence to Support the Uses of the National Teacher Evaluation System (NTES)

I. **The NTES distinguishes among teachers based on effectiveness**
  A. Scoring/rating
    1. The rubrics and scoring procedures are appropriate to each of the components of the NTES (portfolio assessment, supervisor assessments, peer interview, self-assessment) [1a, 1a(p), 1b, 1b(p), 3, 3(p)].
    2. The score on each of the components is reliable (free from rater error). [2, 2(p), 7, 7(p)]
    3. The score on each of the components is fair and free of bias. [2, 2(p), 7, 7(p), 9d]
  B. Generalization
    1. The evidence gathered by each of the components is consistent over occasion, location, and other relevant measurement facets. [8a, 8a(p), 8b, 8b(p)]
  C. Extrapolation
    1. The skills measured by each of the components of the NTES are aligned with the teaching standards (MBE). [1a, 1a(p), 1b, 1b(p)]
    2. The combined score on the NTES appropriately weighs the information gathered by each of the components.
    3. The combined score on the NTES reflects quality of teaching. [4a, 4b, 4c, 4c(p)]
    4. The performance levels were set in an appropriate manner. [6, 6(p)]
    5. The distinctions between outstanding, competent, basic and unsatisfactory teachers are justified. [4a, 4b, 4a(p), 4b(p)]
  D. Consequences
    1. The rewards given to high-performing teachers are fair and appropriate. [5c, 5d, 5e, 5g]
    2. The corrective procedures and dismissal decisions for low-performing teachers are fair, appropriate and administered consistently across municipalities. [5g, 9a, 9b]

II. **The NTES supports teacher improvement and enhances the conditions of teaching**
  A. Scores
    1. NTES scores are valid indicators of teacher effectiveness (See all the evidence listed for I. above)
  B. Reporting
    1. NTES scores are reported in a timely manner.
    2. NTES score reports are clear and contain sufficient information to interpret the scores. [5a(p), 5b]
    3. Teachers, principals and municipal authorities understand the meaning of NTES score reports [5a (p), 5b].
  C. Actions
    1. Professional development and other supports are provided by schools and by municipalities, and teachers participate in these activities. [5a, 5b, 5i, 5j]
    2. Professional development and other supports are linked to needs identified by NTES. [5a, 5b, 5i, 5j]
  D. Consequences
    1. Low-performing teachers' NTES scores and teaching practices improve over time. [5h(p)]
    2. Teachers who earn rewards stay in the profession in increasing numbers. [5e, 5f, 5h]
    3. Teachers whose performance is unsatisfactory repeatedly are not rehired. [5g]
    4. High-performing teachers receive monetary rewards and social recognition. [5c, 5d, 5e]
    5. The NTES fosters collaboration among teachers. [5f]
    6. The NTES improves teachers' job satisfaction. [5c, 5d, 5e]

Note. Numbers in brackets refer to entries in Tables 1 and 2; pending studies are designated by (p).

PENDING ISSUES REGARDING THE VALIDATION OF NTES

Bearing in mind Cronbach's (1989) idea that validation is a "long, even interminable" process (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's 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 coauthors, and by NTES program staff.

Second, we need to investigate the appropriateness of the differentiation between basic and competent teacher performance because these are the categories that distinguish expected from below-acceptable performance, with attached positive versus negative consequences. So far we have examined only the validity of the extreme performance categories (unsatisfactory vs. 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 methodology using longitudinal student achievement data.

LESSONS AND CHALLENGES IN PERFORMING VALIDATION RESEARCH ON HIGH-STAKES ASSESSMENT AND IMPROVEMENT POLICIES SUCH AS NTES

NTES is a high-stakes assessment program that potentially grants salary bonuses to some while threatening with loss of employment to others. It is, at the same time, a formative policy initiative that is meant to improve teaching practices within the classroom and thus reach the long-term goal of increasing student achievement. In this section we discuss the lessons we have learned and the challenges we have faced while conducting validation research of NTES. The characteristics of this program make our experience relevant, we believe, to other researchers facing similar tasks in different settings.

First, its high-stakes nature makes the examination of validity of the assessment, including consequences, highly relevant. High-stakes assessment programs should be considered educational interventions that need to be evaluated like any other educational program (Kane, 2006), and many examples of such research exist in the context of student achievement testing and school accountability in the United States (e.g., Koretz, Barron, Mitchell, & Stecher, 1996; Linn, 2006; Mintrop & Trujillo, 2005; Stecher, Barron, Chun, & Ross, 2000). Examples from teacher evaluation programs, however, are less numerous but increasingly relevant in the United States (for examples, see Milanowski & Heneman, 2001; Odden, 2004).

Considering its relevance, a question in our minds is, What is the optimal timing for validation research? and whether it should start even before the implementation of the assessment system. Although it is not possible to assess all aspects of validity before the implementation of the program because the real conditions in which the program will operate (e.g., scale) and its consequences cannot be fully simulated in pilot settings, initial steps could and should be undertaken early to explore the relationship of the assessment to other variables as well as its internal consistency, reliability, and fairness. It is undeniable that the introduction of high stakes will modify the measurement characteristics of the assessment program (Brennan,

2006), but pilot studies in low-stakes conditions provide an initial indication of assessment characteristics that, in case of not complying with minimum acceptable levels, will inform needed modifications to the instrument design and/or scoring process.

One of the elements that cannot be studied in advance, however, is the "corruptive influence" of high stakes on examinees' behavior and test use (Brennan, 2006) and its effects on the validity of the assessment program as it introduces irrelevant variation in scores. This corruptive influence in NTES has taken the form of "paid portfolios" sold over the Internet and among acquaintances as well as paid consultants who help complete the portfolio. Although there are rumors of these types of behavior, we do not have objective data about how extended the problem may currently be and how it may be altering scores. Brennan (2006) referred to the high cost of preventing such negative behavior: "Protective mechanisms to preclude such misuse are very expensive (e.g., newly developed forms every year) and/or can create an unpleasant educational environment (e.g., intrusive monitoring of school administrators, teachers and students)" (p. 10).

In the case of the NTES, validation research began concurrently with implementation, which has been a challenge because the programs are constantly morphing due to external (nontechnical) issues. Once a high-stakes policy is in place, laws and regulations change over time, sometimes from one year to the next, and that in turn changes the nature of the program as well as its implications and consequences. For example, starting in 2006 and in order to increase teacher participation in NTES, all teachers who refuse to be evaluated automatically receive the "unsatisfactory" performance level, and since 2011 there are more high-stakes consequences attached to below-expected performance. These changes should be taken into account when performing research about the consequences of an assessment system.

Ironically, once the assessment is being implemented it becomes very hard to introduce technical modifications. In the case of NTES, for example, making evidence-based improvements, such as aligning the portfolio and reports to the empirically detected factors, has been avoided as program staff considered it too politically sensitive. Program staff feared that the introduction of changes may cause the system to lose some of its credibility, especially in the eyes of those critical of the system. The need for continuity and consensus, however, needs to be balanced with the findings from validation research. We believe that suggestions from researchers should be taken into account for improving the system and validity research findings should be widely disseminated to provide all stakeholders with information on the strengths and weaknesses of the assessment system.

As already mentioned, NTES was borne out of difficult negotiations between three main stakeholder groups (Ministry of Education, teacher union, municipalities association; Avalos & Assael, 2006), and this makes validation research more complex and politically relevant. The political agreement that was reached between these three parties in 2001 did not reflect the entire membership base of the Teacher Union, and disagreement was shown by a significant proportion of teachers refusing to undergo evaluation during the first years of NTES implementation. This lack of support has decreased over time (while in 2005 about 30% of teachers called to be evaluated that year refused to submit their portfolio, by 2008 that proportion amounted to only 5%, not least due to legal obligations added to the teacher evaluation law). However, changes in the leadership of the Teacher Union have meant a shift away from its initial endorsement of the evaluation toward a more radical stance manifested in a rejection of the evaluation by the head of this union (see Herrera, 2009). There is an advisory board consisting of members

from all the stakeholder groups who meet to review each year's evaluation results and lessons learned, and to make any necessary adjustments to the system. This continued participation of all stakeholders in the design and implemention of NTES ensures a broader consensus for the program but also makes consensus on what should be changed harder to achieve. In such a complex political context, validation research risks becoming politically charged. Therefore the questions of who sets the validity research agenda, who decides how research results are comunicated, and who decides what modifications are suggested based on research results are central to the discussion and future of the program.

In our case, the first two authors worked at the university measurement center responsible for implementing the evaluation. As researchers we were independent from the assessment staff and had autonomy to set our own research agenda (the research area is an independent unit within the center), but at the same time we were aware of the importance of this large-scale assessment program for the center. The reporting of negative findings is a delicate issue as it could be used as ammunition to terminate the program. The closeness to the program's implementers also brought us some benefits: easy access to data and good rapport with knowledgeable staff who can provide detailed information about instrument design and program implementation, and give advice on validation study design and data collection. This relationship is bidirectional as we see our research oriented, in part, to help program staff improve the design and implementation of the program. But we are also interested in communicating our findings to the national and international academic community.

Another important and related issue is who funds validation research. In our case, things have evolved positively over time: Although at first our research was funded by the assessment center budget, later we managed to get external funding from a governmental research agency. From our point of view it is a paramount responsibility of the assessment developer to perform validation research, and we performed large part of our validation research from underneath the same roof as the assessment developer. Therefore our validation results will always be subject to suspicions. Notwithstanding this legitimate criticism, we hope to have contributed importantly to making NTES more valid and relevant to all assessment users. We also hope our experience is helpful to other researchers responsible for validating similar assessment and improvement systems.

REFERENCES

American Educational Research Association. (2005). Teaching teachers: Professional development to improve student achievement. *Research Points*, *3*, 1–4.

American Educational Research Association, American Psychological Association & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: 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*, 254–266. doi:16/j.ijer.2007.02.004

Bell, C., Gitomer, D., McCaffrey, D., Hamre, B., Pianta, R. & Qi, Y. (2012). An argument approach to observation protocol validity. *Educational Assessment, 17*, 62–87.

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

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.

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

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

Brown, T. (2006). *Confirmatory factor analysis for applied research.* New York, NY: Guilford.

Center on Education Policy. (2011). *More to do, but less capacity to do it: States' progress in implementing the recovery act reforms*. Washington, DC: Author. Retrieved from http://cep-dc.org

Correnti, R. & Martinez, J. F. (2012). Conceptual, methodological and policy issues in the study of teaching: Implications for improving instructional practice at scale. *Educational Assessment, 17*, 51–61.

Cortés, F., Taut, S., Santelices, V., & Lagos, M. J. (2011, January). *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* [Teacher professional development and the Professional Development Plans in Chile: Strengths and weaknesses in light of international evidence]. Paper presented at the annual meeting of the Chilean Public Policy Association, Santiago, Chile.

Creswell, J., & Plano Clark, V. (2007). *Designing and conducting mixed methods research.* Thousand Oaks: Sage.

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

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

Gaertner, H., & Pant, A. (2011). How valid are school inspections? Problems and strategies for validating processes and results. *Studies in Educational Evaluation, 37*, 85–93.

Haertel, E. (2006). *Reliability.* In R. Brennan (Ed.), *Educational measurement* (4th ed., pp. 65–110). Westport, CT: Praeger.

Herman, J., & Baker, E. (2009). *Assessment policy: Making sense of the Babel.* In G. Sykes, B. Schneider, & D. Plank (Eds.), *Handbook of education policy research* (pp. 176–190). New York, NY: Routledge.

Herrera, J. (2009, September 22). 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*, p. 17.

Hill, H. C., Kapitula, L., & Umland, K. (2011). A validity argument approach to evaluating teacher value-added scores. *American Educational Research Journal, 48*, 794–831.

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*, 5–35. doi:10.1016/j.apmr.2010.02.005

Kane, M. T. (2006). *Validation.* In R. Brennan (Ed.), *Educational measurement* (pp. 17–64). Westport, CT: Praeger.

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

Koch, M. & DeLuca, C. (2012). Rethinking validation in complex high-stakes assessment contexts. *Assessment in Education: Principles, Policy & Practice, 19*, 99–116.

Koretz, D. (2009). *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.

Leal, P., & Santelices, V. (2010a). *Análisis situación docentes eximidos y suspendidos sistema de Evaluación Docente* [Analysis of the teachert's status as exempt and suspended in the national teacher evaluation system] (Internal Tech. Rep. MIDE UC). Santiago: Pontificia Universidad Católica de 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 Tech. Rep. MIDE UC). Santiago: Pontificia Universidad Católica de Chile.

Linn, R. (2006). *Educational accountability systems* (CSE Tech. Rep. No. 687). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

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.

Manzi, J., González, R., & Sun, Y. (Eds.). (2011). *La evaluación docente en Chile* [The Chilean Teacher Evaluation]. Santiago, Chile: Facultad de Ciencias Sociales, Escuela de Psicología, PUC.

Manzi, J., Strasser, K., San Martin, E. & Contreras, D. (2008). Quality of Education in Chile. Washington, D.C.: Inter-american Development Bank. Retrieved November 8, 2012 from: http://www.iadb.org/en/research-and-data/institution-details,3191.html?id=144

Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York, NY: Macmillan.

Messick, S. (1995). Standards of validity and the validity of standards in performance assessment. *Educational Measurement: Issues and Practice, 14*, 5–8.

Milanowski, A., & Heneman, H., III. (2001). Assessment of teacher reactions to a standards-based teacher evaluation system: A pilot study. *Journal of Personnel Evaluation in Education, 15*, 193–212.

Ministry of Education. (2004). *Marco Para La Buena Enseñanza* [Guidelines for good teaching]. Santiago, Chile: Ministerio de Educación.

Ministry of Education. (2008). Informe nacional de resultados SIMCE 2007 [National report of SIMCE 2007 results]. Santiago, Chile: Unidad de Curriculum y Evaluación.

Ministry of Education. (2009). Informe nacional de resultados SIMCE 2008 [National report of SIMCE 2008 results]. Santiago, Chile: Unidad de Curriculum y Evaluación.

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 L. Ingvarson & J. Hattie (Eds.), *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*, 253–285.

National Council on Measurement in Education. (2010, March). *NCME Newsletter, 18*(1). Available from http://ncme.org/default/assets/File/pdf/newsletter/vol_18_num_1_v2.pdf

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* (M. D. Hakel, J. Anderson Koenig, & S. W. Elliott, Eds.). Washington, DC: The National Academies Press.

Odden, A. (2004). Lessons learned about standards-based teacher evaluation systems. *Peabody Journal of Education, 79*, 126–137.

Pecheone, R. L., & Chung, R. R. (2007). *Summary of validity and reliability studies for the 2003–04 pilot year* (PACT Tech. Rep.). 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

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

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

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

Santelices, M. V., Taut, S., Araya, C., & Manzi, J. (2009, April). *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, San Diego, CA.

Santelices, M. 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] (Tech. Rep. MIDE UC, IT 1007). Retrieved from MIDE UC, Pontificia Universidad Cátolica de Chile website: http://www.mideuc.cl/docs/informes/it1007.pdf

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: Pontificia Universidad Católica de Chile.

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.

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

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

Strauss, A. & Corbin, J. (1990). *Basics of qualitative research. Grounded theory procedures and techniques*. Newbury Park, CA: Sage.

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

Sun, Y., Correa, M., Zapata, Á., & Carrasco, D. (2011). *Resultados: Qué dice la Evaluación Docente acerca de la enseñanza en Chile* [Results: What do the results from the Teacher Evaluation say about teaching in Chile]. In J. Manzi, R. González, & Y. Sun (Eds.), *La evaluación docente en Chile* (pp. 91–135). Santiago: Pontificia Universidad Católica de Chile.

Tabachnick, B. G., & Fidell, L. S. (1996). Principal components and factor analysis. In *Using multivariate statistics* (3rd ed., pp. 635–707). New York, NY: HarperCollins.

Taut, S., & Santelices, V. (2007, April). *Validating the Chilean National Teacher Evaluation System: A comprehensive research agenda*. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

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

Taut, S., Santelices, V., Araya, C. & Manzi, J. (2011a, April). *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, New Orleans, LA.

Taut, S., Santelices, V., Araya, C., & Manzi, J. (2011b). Perceived effects and uses of the national teacher evaluation system in Chilean elementary schools. *Studies in Educational Evaluation, 37*, 218–229. http://dx.doi.org/10.1016/j.stueduc.2011.08.002

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* [Re-evaluation results and employment situation of teachers evaluated by the national teacher evaluation system between 2003 and 2008] (Informe técnico MIDE UC, IT1007). Retrieved from MIDE UC, Pontificia Universidad Católica de Chile website: http://www.mideuc.cl/docs/informes/it1007.pdf

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

Valencia, E., & Taut, S. (2008). *Estudio de dimensionalidad del portafolio de Docentemás 2007* [Dimensionality study of the Docentemás portfolio 2007] (Internal Tech. Rep. MIDE UC). Santiago: 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*, 117–132.