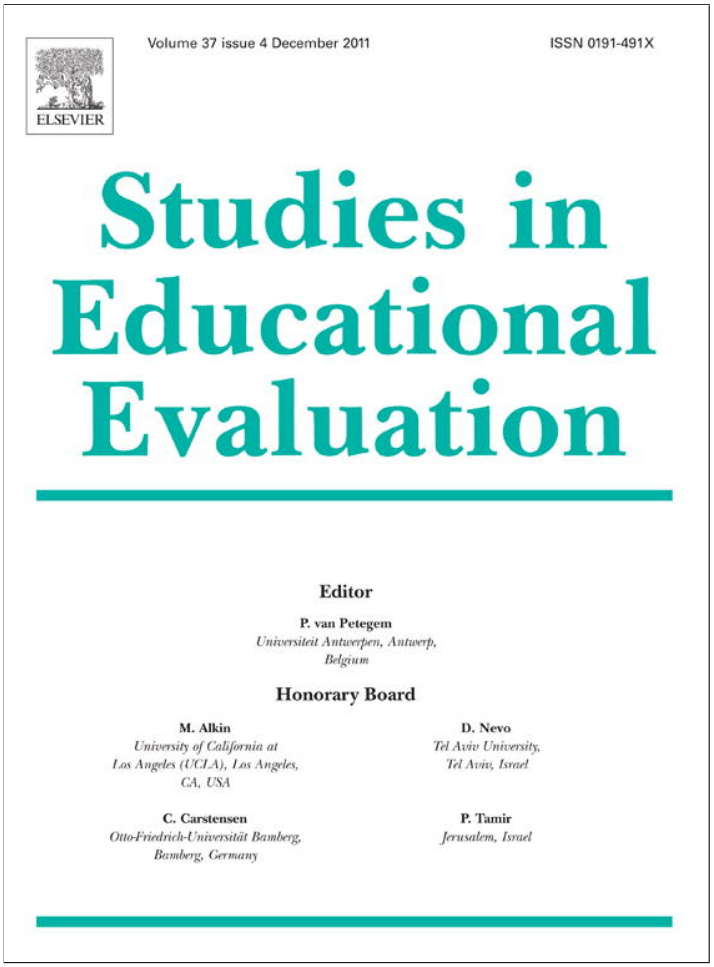

Studies in Educational Evaluation 37 (2011) 218–229

Contents lists available at SciVerse ScienceDirect

Studies in Educational Evaluation

journal homepage: www.elsevier.com/stueduc

Perceived effects and uses of the national teacher evaluation system in Chilean elementary schools

Sandy Taut *, Maria Verónica Santelices, Carolina Araya, Jorge Manzi

Pontificia Universidad Católica de Chile, Chile

ARTICLE INFO

Article history:
Received 4 January 2011
Received in revised form 1 August 2011
Accepted 18 August 2011
Available online 30 November 2011

Keywords:
Consequential validity
Standards-based teacher evaluation
Teacher performance assessment
Consequences of high-stakes assessments in schools

ABSTRACT

This paper addresses the perceived consequences of the Chilean national teacher evaluation system. We interviewed 57 school leaders in 30 schools across 10 municipalities about effects and uses of the assessment in their schools. Results show that in the large majority of schools our interviewees observe positive effects such as increased teamwork and internal reflection processes based on the assessment results. Reports about effects at teacher level far outnumber institutional effects, and are mixed. In all schools our interviewees report on teachers' negative emotions and work overload due to the assessment process, but also about their internalization of the underlying professional standards. The paper analyzes differences among schools and offers suggestions for the development of large-scale standardized teacher evaluation systems.

Introduction

This paper addresses the perceived consequences of the Chilean national teacher evaluation system (NTES) in elementary schools. This study was conceived within the framework of a larger study that addresses the intended and unintended consequences of this assessment system at local (municipal), school and individual (evaluated teacher) levels. In this paper we focus on the perceived effects and uses in elementary schools, as reported by school leaders. We focused on what happened to the school as an institution, to its teachers, and to the school leaders themselves. We also wanted to better understand why these consequences varied between schools. We explore these questions from the perspective of measurement professionals charged with the validation of the NTES, and we conceptualized NTES' consequences as an important type of validity evidence (consequential validity).

Participation in the assessment has been mandatory by law for all public (municipal) school teachers in Chile since 2005. The assessment is based on national standards describing good teaching (Ministry of Education, 2004). Evaluation methods include (1) a portfolio reflecting teaching materials related to a pedagogical unit and a videotaped lesson, (2) supervisor assessment, (3) peer interview, and (4) self-assessment. The system has both a formative purpose and high-stakes consequences for teachers, distinguishing between "outstanding", "competent", "basic", and "unsatisfactory" performance. Outstanding and competent teachers are eligible for an increase in salary after passing a subject knowledge test, while basic and unsatisfactory teachers are subject to mandatory professional development, and – if repeatedly evaluated unsatisfactory – loss of employment.

The Standards for Educational and Psychological Testing indicate that the uses of assessments programs, such as NTES, often involve claims for intended benefits that go even beyond direct uses of the scores, and that these claims should be examined if they are central to the rationale given for implementing the assessment (AERA, APA, & NCME, 1999). The standards refer to the evidence associated with the testing program's consequences as one of the five sources of evidence that should be examined when constructing the unitary validity argument for a given assessment program for a specific use, and according to Sireci (2009), Linn (2009), Lane and Stone (2002) and others, the evidence should include both intended and unintended consequences. The research presented in this article responds to the charge of the standards and explores the perceived consequences of the NTES at the school level as reported by principals and pedagogical experts, with a special emphasis on unintended effects.

Consequences of assessment programs and high-stakes accountability systems

Accountability policies are considered by Ball (2008) part of the "new public management (NPM)" approach to education, a movement that promotes the use of the private sector culture and management procedures in the public sector. According to the

* Corresponding author at: P. Universidad Católica de Chile, Escuela de Psicología, Centro de Medición MIDE UC, Edificio MIDE UC, piso 8, Avda Vicuña Mackenna 4860, Macul, Santiago, Chile. Tel.: +56 2 354 5302; fax: +56 2 354 1729.
E-mail addresses: staut@ucla.edu, staut@uc.cl (S. Taut).

0191-491X/$ – see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.stueduc.2011.08.002

NPM, rationalization and standardization of processes, emphasis on the output and control mechanisms are ways to ensure education quality (p. 48). Ravitch (2010) shares this perspective. Brennan (2006) observes that "most policy makers assume that accountability in education can be accomplished only through the imposition of high stakes testing... When testing becomes high stakes, it is almost inevitable that it will drive instructional decisions, usually by narrowing the curriculum in the direction of emphasizing the content and skills tested. This may be an unintended outcome, but it has real consequences that may not be desirable (p. 10)." Ravitch (2010) concurs, "testing was not the problem. Tests can be designed and used well or badly. The problem was the misuse of testing for high stakes purposes,..." (p. 150).

There is general consensus on the need to document and investigate the consequences of educational measurement and evaluation systems that are used for accountability purposes – whether such studies are conducted as part of the validation process itself or not (Linn, 1997; NCME Newsletter, 2010; Popham, 1997; Shepard, 1997). When conducted as part of the validation of an assessment program, as in our case, the investigation of the consequences is referred to as *consequential validity* (Lane & Stone, 2002; Lane, Park, & Stone, 1998; Messick, 1994, 1998; Moss, 1998; Shepard, 1997). Kane (2002) states that in order to establish the validity of high-stakes testing programs, the underlying assumptions must be examined and criticized (also see Forte Fast & Hebbler, with ASR-CAS Joint Study Group on Validity of Accountability Systems, 2004; Linn & Baker, 2002). The higher the stakes of the testing programs and the stronger the intentions of using them as tools to reform educational institutions, the more the testing program should be considered as an educational intervention. The comprehensive evaluation of educational interventions includes the evaluation of intended and unintended outcomes of the program evaluated (Kane, 2006). Kane says:

> "... For stakeholders to make informed decisions about the effectiveness of high-stake 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 (2006, p. 56)."

The framework proposed by "Joint Study Group on *Validity* in Accountability Systems" emphasizes the implicit program theory, or underlying assumptions, held by designers and implementers as the basis for studying their consequential validity (Forte Fast & Hebbler, with ASR-CAS Joint Study Group on Validity of Accountability Systems, 2004; Kane, 2002; Linn & Baker, 2002; Lane & Stone, 2002). Patton (1997) suggests that the articulation of a program's theory needs to consider that the policy makers are seldom familiar with its details and that, once rebuilt, program theories tend to have conceptual holes that need to be filled (also see Weiss, 1973). Lane et al. (1998) present a methodological framework for evaluating the validity of assessment programs based on the triangulation of different sources of information.

Consequences of high-stakes accountability systems, especially unintended negative ones, have been the interest of researchers in those countries where such policies have been implemented. These researchers have examined the validity of score gains, consequences on instructional practice, teacher and student morale, classroom climate and organizational cohesiveness (Koretz & Hamilton, 2006). Although most of this research has referred to school accountability based on student achievement results, findings may be informative for teacher accountability based on teacher performance assessments such as the NTES.

Koretz and Hamilton (2006), in a review of recent research examining the effects on instructional practices of high-stakes student achievement measurement programs in the United States such as No Child Left Behind, highlight both positive and negative changes in teachers' behaviour. For example, teachers worked harder and focused more on achievement than they had before implementing high-stakes testing, but they also reallocated time to put more emphasis on tested materials and neglected material that was not tested.

Herman and Baker (2006), along the same lines, summarize unintended negative consequences occurring in the United States as result of federal and state accountability programs, such as schools focusing on the test rather than the standards and even ignoring the broader domain of disciplines and subjects that are not tested. As a result students, especially those at-risk, are facing a narrower curriculum of mainly mathematics and reading. The authors also present the case of North Carolina where accountability exacerbated the problem of attracting and retaining quality teachers in low performing schools, the case of the high school exit exam which has increased high-school drop-out rates, and the phenomenon of test score inflation – higher test scores that do not translate into learning gains in other settings.

Given all of these findings and observation based on U.S. high-stakes student testing, the NTES provides an interesting case for studying the consequences of a high-stakes *teacher* assessment system in a context different from the well-studied U.S. American environment.

Description of Chile's high-stakes national teacher evaluation system (NTES)

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

In 2008 Chile had roughly 176,500 classroom teachers, of which 55% worked in municipal schools (Ministry of Education, 2010). Teachers currently do not have to pass a teacher licensure exam that would allow them to start their teaching practice. In municipal schools teacher wages are linked to a state minimum wage, seniority, bonuses for additional training, geographic placement, and managerial responsibility, as well as bonuses that are based on an accreditation of excellence to schools [Sistema Nacional de Evaluación de Desempeño Profesional, SNED], and an individual certification of excellence [Asignación de Excelencia Pedagógica, AEP].

The national teacher evaluation system (NTES) was introduced by the Ministry of Education in 2003, and since 2005 is mandatory for teachers in municipal schools nation-wide. Performance standards guiding the evaluation have been defined, officially endorsed, published and widely disseminated as the "Guidelines for Good Teaching [Marco Para la Buena Enseñanza]" (Ministry of Education, 2004; Ministry of Education Legal Department, 2004). The NTES is the result of an agreement between three main stakeholder groups that traditionally hold opposing political views and to this day have both shared and diverging expectations regarding the program. The NTES 2009 results show that the majority (63.1%) of evaluated teachers received the performance categorization of "competent", while 28.9% were evaluated as showing "basic" performance. Only 6.5% were evaluated as "outstanding", and a mere 1.5% were considered as "unsatisfactory". Similar distributions of NTES results were observed in previous

years. For more details on the development of the teacher evaluation system and its characteristics see Avalos and Assael (2007), or Manzi, González, and Sun (2011).

The NTES is a mandatory, high-stakes evaluation system used to reward and sanction public school teachers: those teachers who are found to be "competent" or "outstanding" are eligible for an increase in salary,[1] while basic and unsatisfactory teachers are subject to mandatory professional development, and – if evaluated "unsatisfactory" in two consecutive years – loss of employment.[2] Teachers showing "basic" performance – the majority of evaluated teachers so far – must be reevaluated 2 years later and they get three chances to improve their performance to the expected performance level of "competent" before being subject to termination. At the same time, the NTES' formative purpose finds expression in the associated professional development courses (called *Planes de Superación Profesional*, or short PSP). They are defined as a set of actions aimed at improving the weaknesses of teachers evaluated as "basic" or "unsatisfactory" by NTES.[3] Since 2005 each municipal educational authority is responsible, and receives funding, for designing and implementing the professional development courses based on their yearly municipal NTES report and, according to the rules that govern the NTES, the design needs to consider the knowledge, competencies, abilities, domains and criteria defined in the professional teaching standards (Ministry of Education, 2004).

Another important formative aspect of the NTES is the descriptive report each teacher receives with the results for the different instruments and portfolio dimensions, describing strengths and weaknesses, and including their final score. The school principal and the head of the municipal education authority also receive reports providing the final performance categories of the teachers evaluated in that school or municipality, and the average results for all those teachers by instrument and portfolio dimension.

Evaluation methods include (1) portfolio assessment comprising a written part and a videotaped lesson, (2) supervisor assessment, (3) peer interview, and (4) self-assessment. The portfolio asks the teachers to describe planning and student assessment materials for a specific, pre-defined set of lessons, as well as to reflect on their use in the classroom. One lesson (45 min) of each teacher is videotaped by an external contractor. Two supervisors (generally the principal of the school and the school's pedagogical expert) complete an evaluation questionnaire asking about professional qualities of the evaluated teacher. The peer interview is performed by another teacher (not from the same school, but teaching the same subject and grade level) based on a structured interview protocol containing questions about pedagogical knowledge and practice. Finally, the self-assessment is a questionnaire that asks the teacher to critically reflect on his or her professional performance.

[1] This incentive program is called Individual Performance Bonus [Asignación Variable por Desempeño Individual, AVDI]. Teachers who perform at the "competent" or "outstanding" level are eligible to apply for a salary bonus if they perform sufficiently well on a test of disciplinary and pedagogical knowledge. The bonus ranges between 5% and 25% of the gross monthly salary and the amount awarded depends on the performance on the test as well as on NTES.

[2] 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.

[3] The professional development courses were created by law (Law No. 19.961) and are part of the special norms that regulate NTES (decree 192). Teachers, however, are not paid for the additional time they need to attend these trainings.

The national teacher evaluation system's goals and intended consequences

The first step in our overarching study of the national teacher evaluation system's consequences was to explicate the underlying program theory regarding NTES' intended effects and functioning. To this end we consulted legal and policy documents as well as prominent stakeholders pertaining to the Ministry of Education, the Teacher Union, the Association of Local Governments, and the implementing agency, all of which were involved in the design of this assessment policy. The program theory explication process brought to light intended consequences at local and individual levels, but there was no explicit mention of effects and uses expected at school level.

At local level, the intended consequences of the assessment that were mentioned by at least two stakeholder groups were the following: (1) offering social reinforcement of good teaching practices, (2) building the capacity of teachers with shortcomings as diagnosed by the assessment, (3) informing the selection and exit of teachers, (4) providing a base for peer conversations about good practice, (5) improving teachers' job prospects by providing access to monetary incentives, (6) informing educational management decisions. At individual level they were: (1) triggering changes in weak practices as diagnosed by the assessment, and (2) maintaining good practices by triggering internal reinforcement of diagnosed strengths (for more details see Taut, Santelices, Araya, & Manzi, 2010).

Despite the lack of explicit expectations at school level, reports of teachers' assessment results are sent to all schools where teachers were evaluated in any given year, which implies expected use of this information at school level. Furthermore, school leaders play evaluator roles in the assessment system, and school is the place where teachers work on their portfolios, have one lesson video-taped, and receive a peer from a different school who conducts the peer interview – school is where the assessment actually "happens." Finally, we think that some of the effects that policymakers expected to happen at local level may be understood as happening in schools as well. For example, both local educational authorities as well as school leaders could offer recognition of good teaching practice based on the evaluation results. Peer collaboration and conversations about good practice seem most likely to happen in schools, and to lesser extent in the local environment. Thus, although policymakers and documents do not explicitly mention intended consequences at school level, for the reasons given above we decided to include this level in our overall study.

Prior to the current study we empirically investigated NTES' consequences at local (municipal) level. We found some intended (positive) consequences, such as installing local reward mechanisms for teachers based on the assessment results, supporting teachers who show shortcomings as diagnosed by the assessment through targeted professional development, and using the assessment results for educational planning (e.g., assigning good teachers to schools in most need). At the same time, the municipalities (with one exception) did not use the assessment information to inform hiring decisions, and none reported using them to inform firing decisions. In all 10 municipalities the local education authorities reported observing unintended consequences, most prominently negative emotions and resistance on the part of teachers, which, however, diminished over time. They also talked about the work overload the installation of the assessment has meant, not only for teachers but also for themselves (for more details see Santelices, Taut, Araya, & Manzi, 2009). Furthermore, the study found that although municipalities coordinated with schools the implementation of the assessment, there was no indication (with one exception) that municipalities

worked systematically with schools on promoting intended effects and giving concrete uses to the assessment process and results.

Previous findings about the effects of standards-based teacher evaluation systems on teachers and schools

Teacher evaluation systems share the common goals of improving instruction and student learning, regardless of the structure of the assessment system (whether it is an internal or external process), who the appraisers are (school administrators, peers, self or external body), and what sources of data are used (peer interview, self-evaluation, classroom observation) (Porter, Youngs, & Odden, 2001; Peterson, 2000).

Recently there has been renewed interest in teacher evaluation systems in the United States, partly triggered by federal legislation. Some of the systems in place now are somewhat similar to the Chilean system in that they combine different data sources, are based on teaching standards delineating expected professional practice, and serve both formative and summative purposes, although few have attached such high stakes consequences to its results as in the Chilean case. Some research deals with the consequences of standards-based teacher evaluation systems on teachers (Heneman & Milanowski, 2003; Heneman, Milanowski, Kimball, & Odden, 2006) while others have focused on its impacts at school level (Halverson, Kelley, & Kimball, 2004; Kimball, 2002), and findings show mixed results at both policy levels. These findings are detailed below.

Overall teachers tend to agree with the competency models reflected in the standards-based assessments under study. Teachers perceive the systems as fair largely because they were based on common, explicit standards, multiple data sources, and gave teachers opportunities for input into the evaluation process (Heneman et al., 2006).

Teachers also describe a positive, yet sometimes transitory, impact on instructional practice. For example, when comparing the implementation of teacher evaluation systems in four different counties, Heneman III and colleagues found changes in instructional practices such as engaging in more reflection, becoming more organized, and improving lesson planning and classroom management (Heneman & Milanowski, 2003; Heneman et al., 2006). Lustick and Sykes (2006) investigated the effect of the U.S. American National Board for Professional Teaching Standards (NBPTS) certification process by using a quasi-experimental design, complemented by in-depth interviews. They concluded that there were two NBPT standards in particular that science teachers learned most about due to the assessment process: using the scientific inquiry method and classroom assessment. They attributed these positive effects mainly to teachers' work with the standards themselves as the "critical carrier for the knowledge base of teaching" (p. 29). Sato, Wei, and Darling-Hammond (2008) also found that teachers who undertook National Board certification changed their teaching practice, particularly their classroom assessment practice, significantly more over the course of their certification than a comparison group of non-participating teachers, and these changes were sustained the year following the certification process.

The literature also describes significant obstacles to the evaluation efforts, such as principals' and teachers' workload, insufficient training for the evaluators, poor evaluation design (Heneman et al., 2006), as well as inconsistent evaluation data gathering and evaluator decisions (Kimball, 2002). Heneman et al. (2006) found that although teachers accepted the standards-based evaluation as "appropriate" and as an "adequate description of good teaching", they "were not enthusiastic about portfolios. These were perceived as burdensome and the requirements confusing." Teachers also described the system as a whole as "disruptive and stressful" (p. 16).

While survey data indicates that a large majority of teachers found their NBPTS certification to have opened up new leadership activities for them (such as promotions to principal or pedagogical-expert positions, or being hired as pedagogical consultants) (Darling-Hammond, Wei, & Johnson, 2009), a more in-depth study on NBPTS certified teachers' impact in schools found that an important barrier to increasing their leadership involvement was the reluctance of principals to award them more responsibilities – be it because of lack of knowledge about NBPTS, skepticism about its value, or a comfort with traditional power relations (Koppich, Humphrey, & Hough, 2006).

The school culture is also an important barrier to accredited teachers' impact in their schools and districts. Koppich et al. (2006, p. 17) describe teachers' adherence to a "culture of egalitarianism" where those who step outside their expected roles and responsibilities can expect some type of colleague rebuke. The authors conclude that principals play a crucial role in ensuring the utilization of certified teachers as a school resource. They must be creative to reorganize their schools to allow teachers time to work together and must allow certified teachers to take on new leadership roles. This is more likely to happen if principals are knowledgeable about the certification process and underlying standards, or are even certified themselves, and in general if they focused their responsibility on teaching and learning instead of on administrative tasks. Furthermore, the culture of the school would have to allow teaching to be a much more public activity so that peer collaboration could be fruitful.

Furthermore, school administrators play a key role in shaping teacher perceptions (Heneman et al., 2006) and brokering the implementation, use and consequences of the evaluation system. Studies show that their capacity to structure more frequent interaction with teachers and align school goals with those of the evaluation system, as well as engaging actively in pedagogical and instructional matters, make the difference between effective and ineffective evaluation systems. Halverson et al. (2004) examined how principals adapted standards-based evaluation systems to fit their school context in order to ensure more beneficial effects. Their study concluded that this adaptation process largely depended on the "principals' self-perception of their role as a leaders and the knowledge and skills they bring to that role, prior evaluation practices in the school and district, and school context factors such as teacher morale and existing challenges facing the school" (p. 39). The authors hypothesized that beneficial use of the evaluation depended on instances for teachers and leaders to interact around instruction, and a common language to facilitate these interactions, and the authors thought this could be facilitated by the teaching standards underlying the evaluation. Kimball's (2002) work in three school districts with newly implemented standards-based teacher evaluation systems showed that important enabling conditions for beneficial use of evaluation feedback included principals playing the role of strong instructional leaders. These principals provided opportunities for teachers to work collaboratively on instructional strategies and identified resources to help teachers further develop their knowledge and skills. Furthermore, principals provide coherence when performance culture changes, linking school goals with professional development, monetary incentives (Heneman et al., 2006) and public recognition (National Research Council, 2008) ensuring positive impact on instructional practice.

Regarding principals' role within the evaluation system, it is important to differentiate programs where principals play the role of evaluators from those in which principals only play a supportive role in teachers' preparation for the assessment. As reported by Setliff (1989), when principals are also the appraisers, the school

climate is negatively affected if there is not an equivalent protective school culture that embraces teacher performance assessment and school accountability.

Positive effects are more likely in schools led by principals who are interested and concerned about their teachers, agree with the idea of accountability at the school level, and worry about the teaching and learning experience of their students (Halverson et al., 2004; National Research Council, 2008; Setliff, 1989). If there is a lack of commitment by teachers or their principals, the odds of achieving the intended consequences of a standards-based teacher evaluation system are diminished and unintended effects are likely to arise.

Research questions

In this paper we report on our research about the perceived effects of the Chilean national teacher evaluation system (NTES) and uses of the assessment results in elementary schools, as reported by the school principals and pedagogical experts. Our investigation focused on their accounts about what happened to the school as an institution, to its teachers, and to the school leaders themselves. We also studied why perceived effects and uses varied between schools.

Methods

Sample

Our sample contained 30 public (municipal) schools from 10 municipalities (approx. three schools per municipality). These 10 municipalities were the same in which we examined municipal level consequences (6 urban in the capital region, 2 urban outside the capital region, and 2 rural). These municipalities were purposively sampled with the intention to represent rural and urban as well as high-, medium- and low-poverty communities.

The selection criteria for choosing schools within our 10 municipalities included (a) sufficient number of evaluated teachers, and (b) student achievement results as measured by the national standardized test (SIMCE), selecting in each municipality one high-performing school, one low-performing school, and one school in the middle. Within each school we scheduled interviews with the principal and the head of the so-called technical-pedagogical unit (from here on we refer to them as "school leaders").[4] In total we conducted $N = 57$ interviews, on average $N = 2$ interviews per school. Participation was voluntary and we did not offer a financial or other type of incentive.

Data collection method and instrument

We conducted semi-structured 1-h personal interviews with school leaders in their schools. All interviewees signed an informed consent. Interviews were tape-recorded. The interview protocol aimed at capturing the consequences, effects and uses of the assessment policy as perceived by the interviewees. We started asking a broad question about what the teacher evaluation process had been like in their school, and what were the effects they had observed, and only later did we probe regarding more specific effects at institutional and teacher levels, as well as regarding their own practice. We also distinguished the assessment process from the assessment results, asking interviewees to describe examples of specific uses of either process or results, if any. Finally, we also wanted to know how school leaders viewed the professional development that was offered to low-performing teachers at local level, and whether there was any recognition for high-achieving teachers at their school.

Data analysis

We transcribed all 57 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 (Hsieh & Shannon, 2005; Miles & Huberman, 1994; Patton, 2002). All interviews were double-coded, resulting in one consented version of each coded interview.

In analyzing the data we distinguished between a code's frequency, presence/absence, and salience. Frequency refers to a count of how many times the code was used in the coding of an interview. Presence/absence refers to whether a code was mentioned at least once (presence), or never (absence), in an interview. Salience refers to whether the code was mentioned in response to the first broad question of the interview. The first question of the interview asked about the consequences the evaluation has had in the school.

We summarized the evidence from each school, both quantitatively and qualitatively, in a *school report*. The reports contained a descriptive part focusing on the reported effects and uses, and an explanatory part where we hypothesized as to why these effects and uses were or were not observed in this school, and discussing (and if possible, resolving) any contradictions between the two interviewees.

We also developed summary tables by school. We distinguished between perceived effects (which could be positive, negative or neutral) and active, concrete uses (only positive). Uses require activities to be implemented, while effects can be happening without active involvement of school actors. We also identified a number of explanatory concepts, which came out of the interviewees' response and – based on the relations we established in our analysis – could either positively or negatively impact the observed effects and uses. We named them "context factors" and "mediating factors". A priori we characterized them as either positive or negative, based on our own judgment.

Finally, we grouped schools according to their level of use and predominant effects, relating effects and uses to possible explanatory concepts, in order to try to better understand observed differences among schools. We developed graphical displays to help us in identifying such patterns (see Miles & Huberman, 1994). Specifically, we looked at whether higher presence of positive explanatory concepts was observed in "top" schools versus "bottom" schools, or vice versa for negative concepts.

Results

What kinds of effects and uses did school leaders talk about?

The analysis of the interview data resulted in 205 sub-codes, structured into 34 code families. Post-coding we further summarized these code families into seven groups according to our research questions (see Table 1): (1) effects at school level, referring to how the school as an institution reacted to the assessment process and results, (2) effects on the school leaders themselves, independent of what happened at institutional level, (3) effects at teacher level, referring to consequences that affected teachers directly, independent of institutional reactions, (4) uses at school (institutional) level (as implemented by school leaders), (5) uses at teacher level (as enacted by teachers, independent of uses at school level), (6) mediating factors, and (7) context factors. We further distinguished positive from negative effects, and we also

[4] The head of the pedagogical unit (UTP) usually guides and supervises all teachers in pedagogical matters (we call them "pedagogical experts"), whereas the principal usually plays an administrative leadership role.

Table 1
Code groups and example codes.

Group of codes	Type	# Of codes within group	Example codes within each group of codes
1. Effects at school (institutional) level	Positive	7	– Fostering team work at institutional level
	Negative	2	– Negative effect on work climate
	No effect	2	– Explicit mention of no effect
2. Effects on school leaders	Positive	3	– Learning due to the assessment
	Negative	2	– Teachers blame school leaders for their bad results
	No effect	2	– Explicit mention of no effect
3. Effects at teacher level	Positive	15	– Results triggered revisions of teaching practice
	Negative	14	– Increased workload
	Neutral	4	– Resistance toward NTES has evolved positively
4. Uses at school (institutional) level	Positive	11	– Diagnosis of school's teaching quality
	No use	3	– Limited use or no use (explicit)
5. Uses at teacher level	Positive	2	– Sharing contents of individual assessment report
	No use	1	– Teachers only see final score without taking into account feedback provided by report
6. Mediating factors	Positive	25	– Legitimacy awarded to the assessment program
	Negative	63	– Lack of consistency between assessment results and own perceptions
7. Context factors	Positive	5	– Supportive (pedagogical) role school leaders play in school independent of assessment
	Negative	3	– Lack of self-criticism of teachers as a professional group
	Neutral	16	– Work experience of school leaders

recorded the explicit mention of lack of effects. In the case of concrete uses we found positive uses and lack of use. We also distinguished positive and negative explanatory concepts (see Table 1).

Next we present the aggregate results for the above-mentioned seven types of consequences. This will provide an answer to the question at what level our interviewees perceived most effects and uses, and whether these were mostly positive or negative in nature.

At what level do school leaders perceive the most prevalent effects and uses, and are these mostly positive or negative?

We found that there were many more responses about teacher-level effects than about school-level effects, and very little at school leader level. This is reflected not only in the number of codes related to each of these three groups, and the frequency and the presence of these codes, but also in the salience of these issues. That is, teacher-level issues tended to come up a lot in the first broad question regarding observed effects, meaning they were most top-of-mind (or salient) for our interviewees. Furthermore, at the teacher level interviewees talked more about negative than positive effects, in contrast to the school level where they commented on more positive than negative effects.

Active uses at school level were much more present in interviewees' responses than uses by teachers. Most of the mentions of teacher-level uses concerned the observation that teachers shared their assessment results among each other, which seems to be not a use in itself but rather a prerequisite to active use of the information by teachers as well as schools. Therefore, in what follows we do not report in more detail on teacher-level uses.

In summary, contrary to the program theory underlying the assessment system, which did not explicitly incorporate any consequences at school (institutional) level, we find that some effects and a lot of uses were reported at this level, and these were mainly positive in nature. However, what interviewees talked about most were teacher-level effects, and these were overall mixed.

Next we present more detailed findings regarding each group of codes. We first present in more detail the observed *effects* at school, school leader and teacher levels, followed by details regarding the concrete *uses* that interviewees reported at institutional (school) level.

What effects do school leaders perceive at their schools at an institutional level?

Overall, NTES' effects at institutional level tended to be positive (see Table 2). Nevertheless, there was one particularly prevalent negative effect: interviewees from two-thirds of the schools ($N = 23$ out of 30 schools) mentioned that they had to allocate time for teachers to work on the NTES requirements, which was evaluated as a negative effect by the school leaders because it took time away from other activities.

On the positive side, if NTES results were positive, then in about two-thirds of the schools ($N = 19$) this served as external validation for the good work done by school leadership and teachers. In addition, in about half of our schools ($N = 16$) NTES fostered teamwork between teachers on the one hand and school leaders on the other hand, describing an effect in terms of the institutional work style, instead of an effect in terms of collaboration among peers (the latter is categorized as a teacher-level effect). In a few schools other interesting positive effects included increased rapport between school actors at different levels of authority ($N = 7$) and promoting the installation of an internal evaluation system ($N = 6$).

In terms of NTES' effect on work climate, we found somewhat mixed results: in two-thirds of schools ($N = 19$) there was a negative effect because of elevated stress levels due to the assessment process or due to feelings of injustice among colleagues receiving NTES results different from one another. However, in one third of schools our interviewees reported positive ($N = 11$) or no effects ($N = 9$) in this regard.

What effects do school leaders perceive for themselves?

About a third of principals as well as curricular experts we interviewed explicitly said that NTES had no effect on them personally. In a few cases NTES resulted in learning on the part of the principal and the curricular expert, for example, about what is seen as good teaching practice today. In a handful of schools the evaluation had negative effects on school leaders because teachers

Table 2
Summary of most common effects and uses for schools and teachers.[a]

Positive	Negative
School level effects and uses	
Schools engage in reflection of NTES results (20/30)	Schools have to allocate time to work on NTES (23/30)
NTES provides external validation for good work by school leaders and teachers (19/30)	NTES has negatively affected school climate (19/30)
Teachers with good NTES results receive informal recognition in their schools (18/30)	
NTES has fostered team work at an institutional level (16/30)	
NTES results are used to diagnose teacher performance in the school (16/30)	
Teacher level effects and uses	
NTES has increased peer collaboration related to the evaluation itself (28/30)[b]	*Teachers experience negative emotions (stress, anxiety) during the evaluation process (29/30)*
NTES improves teaching by stimulating its revision through reflection (25/30)	Teachers experience negative emotions (disappointment, anger) due to their bad assessment results (26/30)
Teachers experience positive emotions due to their good assessment results (21/30)	*NTES means excessive workload for teachers (26/30)*
Initial resistance toward NTES has diminished over time (20/30)	Due to the lack of knowledge about NTES teachers resist to be evaluated (19/30)
	General resistance to be evaluated (18/30)
	Others (peers, parents, superiors) devalue teachers due to their bad assessment results (15/30)

[a] Shows effects and uses reported in at least half of all schools (15 out of 30).
[b] Effects that were salient in at least one fourth of all schools (8 out of 30) are shown in italics.

either held them responsible for their bad results, or accused them of injustice when serving their role as evaluators (i.e., giving preferential treatment to some teachers on the basis of personal relationships). Overall, based on their self-perception NTES did not seem to have importantly influenced school leaders.

What effects do school leaders perceive their teachers to experience?

We found mixed results in terms of the assessment system's effects at teacher level (see summary in Table 2). Overall, interviewees talked about as many positive as negative kinds of effects. In almost all schools ($N = 29$ out of 30 schools) our interviewees observed teachers who showed negative emotions (stress, anxiety) during the assessment process and as a consequence of weak assessment results ($N = 26$). If assessment results were satisfactory, however, then in two thirds of the schools ($N = 21$) this caused positive emotional reactions on the part of teachers. Positive emotions (for example, a sense of challenge) were also present in some schools during the assessment process ($N = 11$). School leaders also frequently noted that teachers were devalued by others (peers, parents, the public) if they received bad assessment results ($N = 15$), to much larger extent than being valued for good results ($N = 8$). Also prevalent was teachers' work overload due to the assessment tasks ($N = 26$), which apparently caused them to neglect their other duties during the assessment process ($N = 11$).

There were also numerous positive effects at teacher level. For example, in virtually all schools ($N = 28$) teachers' collaboration (as collaboration among peers, specifically, and not involving school leaders or staff) was strengthened by uniting them in developing the evaluation instruments, observing each other in the classroom, or helping each other interpret the evaluation report. Furthermore, according to our interviewees in 25 out of 30 schools the assessment prompted teachers to revise and reflect on their teaching practice, especially regarding lesson planning and classroom assessment practices. In this revision process the standards underlying the assessment (known as "Marco para la Buena Enseñanza", short "MBE") played an important role: in 13 out of 30 schools our interviewees attributed improvements in teaching to the fact that teachers were forced to review these standards.

Interviewees at nine schools said that improved teaching practice was due to the comparisons between peers that the assessment enables, and in eight schools interviewees noted that simply holding teachers accountable produced such improvement, because the assessment "professionalized" teaching ($N = 8$). However, there were also a number of interviewees who were skeptical that the assessment triggered any improvements in teaching practice (in $N = 13$ schools). Teachers' participation in the economic incentives program was another positive effect that school leaders mentioned in one third of schools ($N = 10$).

In terms of negative effects, in two-thirds of participating schools ($N = 18$) teachers' resistance toward the evaluation system was an interview topic. School leaders attributed the main cause of this resistance to teachers' lack of knowledge about the evaluation process ($N = 19$). Because this knowledge has increased over time it makes sense that in an important number of schools ($N = 20$) teachers' resistance was reported to have diminished since the installation of the evaluation policy. On the other hand, in a few schools the issue of outright refusal ("rebellion") of teachers to participate in the evaluation was still commented on (in $N = 7$ schools), and reasons were analyzed (i.e., their age and teacher union participation). Interviewees also reported that teachers looked for legal loopholes to avoid the evaluation (in $N = 6$ schools).

What uses do schools give to the assessment process and results?

As use we understand a concrete action triggered by the assessment undertaken by school leaders and implemented at institutional level. In two-thirds of our schools the interviewees commented that there were instances when school actors reflected on the assessment results (in $N = 20$ out of 30 schools). Also common seems using the assessment for diagnosing the teaching quality of their staff ($N = 16$).

In terms of recognition practices, almost two-thirds of schools ($N = 18$) informally recognized and rewarded good assessment results obtained by their teachers, for example, by congratulating them in the hallway or mentioning the results in the staff meeting. About half of the schools ($N = 13$) had more formal recognition practices in place (often in addition to informal ones), for example, a ceremony, a gift, or a letter to parents.

At first glance it seems contradictory that in about half of the schools we visited ($N = 14$) at least one interviewee said at first that there was limited or no active use of the assessment in their schools, at an institutional level. Often our interviewees started out responding there were no concrete uses, but upon further thought did think of things that were done in their schools that could (more indirectly) be linked to the assessment. In general, uses were less explicit, conscious, or "top of the mind." Oftentimes we noted uses when analyzing interviewees' responses but when asking them directly they often struggled to pinpoint them.

Some other uses were mentioned in a few schools, among them to use teachers with good assessment results as mentors or leaders in the schools (in $N = 8$ schools), assigning them to difficult or important classes or grades (in $N = 5$ schools), and for marketing of the school (in $N = 5$ schools). Six schools organized their own professional development in order to remedy weaknesses diagnosed by the assessment.

What explanatory variables relating to effects and uses did we identify in the interviews?

As described above, we analyzed each interview not only identifying effects and uses of the NTES, but also looking for possible explanations for strong use or lack of use, or positive versus negative effects, in each school. First of all, this analysis resulted in what we call mediating factors: concepts we think might mediate NTES' effects and uses, both in a positive or a negative direction.

On the negative side, at least one of the interviewees in almost all schools ($N = 28$) stated that the evaluation lacked legitimacy in their eyes, due to shortcomings in its design, instruments, or implementation. The interviewees also reported their teachers to perceive such lack of legitimacy, although to somewhat lesser extent (in $N = 22$ schools). Another important issue mentioned in a large majority of schools ($N = 28$) was the lack of consistency between NTES results and results obtained using other criteria to judge teacher performance. This means that in most schools other criteria than those reflected by NTES were used to judge teacher performance, for example, student achievement, parent relations, commitment to school improvement, or experience. Some school leaders also mentioned other obstacles to using the NTES results, for example, that teachers were not willing to share their NTES results openly with colleagues, or the lack of detail in NTES evaluation reports.

On the positive side, in almost all schools ($N = 28$) at least one of the school leaders described playing a significant role regarding the evaluation, facilitating its implementation and utilization. For example, they motivated teachers to do well on the evaluation, provided psychological and pedagogical support during the assessment process, and helped teachers interpret the evaluation results. It was much more common that the technical-pedagogical supervisor played a pedagogical support role during the assessment (in $N = 24$ schools) than the school principal (in $N = 12$ schools). Interestingly, most school leaders did not mention their new roles when we asked them what effects the evaluation had caused on their own practice. Instead, these roles were mentioned when talking about how they had faced the installation of the assessment system in their schools and how they encountered the consequences this brought about for their teachers, and therefore we categorized these roles as mediating factors for teacher-level effects, instead of as direct effects on school leaders themselves.

In almost all schools ($N = 28$) at least one of the interviewees talked about the observed consistency of results their teachers obtained in the assessment, compared to criteria or impressions of their own. Thus, the assessment both confirmed impressions, as well as provided surprises for school leader regarding their teachers' performance, and these surprises sometimes undermined the evaluation's credibility while at other times school leaders accepted the external assessment result. Some school leaders talked in positive terms about the legitimacy of the evaluation system, both in their own eyes (in $N = 14$ schools) as well as in the eyes of their teachers (in $N = 6$ schools). Oftentimes, the same interviewees legitimized one aspect of the assessment (e.g., the portfolio instrument) while questioning the legitimacy of another (e.g., its punitive consequences). A facilitator to assessment use mentioned in two thirds of the schools ($N = 20$) was that teachers were willing to share their NTES results with colleagues.

We also recorded a number of what we call context factors, as mentioned by the interviewees. These helped us better understand and characterize each school's context. These issues were mainly related to a school's socioeconomic composition, student achievement, working climate, teachers' professional culture, existence of an internal evaluation system, and attitudinal aspects of the interviewees we found relevant but not directly related to the evaluation itself (e.g., their attitude toward evaluation generally, not NTES specifically). This information was used in the explanatory analysis we describe next.

Exploring explanations for differential school-level effects and uses

In order to better understand why there were schools that reported a lot of positive school-level effects and active uses on the one hand, and on the other hand there were schools that reported few positive effects and few uses, we worked on building groups of similar cases. We used both quantitative and qualitative criteria to help us in this task. When defining the quantitative and qualitative criteria the idea was to have three groups of about equal number of schools distinguishing between strong positive effects and uses, medium or mixed effects and uses, and weak positive effects and uses.

We looked at overall frequency of positive school-level effects and uses in each school. Looking at the entire sample, we arbitrarily defined "top cases" as those with a total frequency of positive effects larger or equal 5 and uses larger or equal 6, and "bottom cases" as those with a total frequency of positive effects smaller or equal 1 and uses smaller or equal 2. We also read the school reports we had written and scored each school as either 0 = low, 1 = medium, or 2 = high, depending on the level of use and positive effects described at school level. Again, we wanted to build groups that were internally consistent, i.e., similar among each other, but different from the adjacent group. Three different judges performed this analysis, and we took as the final score the one on which at least two judges had agreed. At the end we combined the qualitative and quantitative judgments: we decided to be very selective and included cases in the top and bottom group only if both quantitative and qualitative criteria came to the same conclusion, leaving the "unsure" cases for the "medium" group. We thus formed a "top group" containing $N = 7$ schools and a "bottom group" containing $N = 9$ schools.

It is interesting to see what kinds of differential effects and uses serve to confirm these two groups of schools. For example, the large majority of schools reported adapting to the evaluation by allowing teachers to allocate working time to the process; of those seven schools not reporting this adaptive reaction, five are in the "bottom" group. On the other hand, of the eight schools reporting to use high-performing teachers as an additional resource, four are found in the "top" group and the other four are in the middle group, with none of the "bottom" schools showing this type of use.

The next step in our explanatory analysis was to see whether some of the mediating factors and context factors we had identified based on interviewees' responses could explain the "top" versus "bottom" grouping of schools. We did this by looking

at the presence of positive and negative mediating factors in "top" versus "bottom" groups. Were positive factors present to higher extent in "top" schools? Were negative factors less prevalent in "top" schools as compared to "bottom" schools? While we could not identify any pattern when analyzing the impact of our context factors, we did find that some mediating factors seemed to show the expected pattern. Particularly, we found that the difference in the presence of positive mediating factors in general was especially large between "top" and "bottom" schools. As for the negative mediating factors, they were in fact more prevalent in "bottom" schools, but the overall difference between the groups was not as striking as the difference we found regarding positive mediating factors.

Based on further qualitative analysis, three enabling mediating factors stood out. First, in schools with positive effects and active uses our interviewees tended to legitimize the evaluation, emphasizing its utility and technical adequacy. It is striking that for the "bottom" schools such legitimization was almost completely lacking. However, in both groups there were also responses questioning the legitimacy of the assessment system. The difference was that in "top" schools, school leaders tended to be more balanced and mentioned both negative and positive aspects of the evaluation, also indicating a more profound knowledge of the details of the assessment.

Second, school leaders in the "top" schools played an active role regarding the teacher assessment, especially in pedagogically supporting their teachers during the assessment process, i.e., for example, helping them prepare the assessment portfolio. When looking at this in more detail, we found that there was a striking difference in the active role in particular the school principal played in this regard in the "top" cases (in six out of seven "top" schools the principal gave pedagogical advice, versus in none of the "bottom" schools). Important here seems to be not only whether he or she became involved in the assessment process, but also whether he or she generally played a pedagogical support role in the school, independent of the assessment itself (in five out of seven "top" schools this was the case, versus in two out of nine "bottom" schools).

Third, in contrast to "bottom" schools, and as a prerequisite for active use, teachers in the "top" schools openly shared and discussed their assessment results. This was not related to whether assessment results were found to be good or bad, but instead could be an indication of a positive, trusting work climate or culture in these schools, which is important for constructive use of performance assessments.

Discussion

This investigation reports on consequences of the Chilean national teacher evaluation system (NTES) taking place in a particular setting: the school, as perceived by school leaders. Although the reports from our interviewees may differ from the "actual" impact of NTES, we consider their accounts to be valid information about the extent to which the policy is achieving positive or negative consequences in schools. The perceptions of principals and pedagogical experts are particularly important since our study was exploratory: school level consequences were not part of the explicit NTES program theory. Linn (2009) recommends exploring especially unintended consequences of measurement or educational programs using qualitative research, hence the methods chosen to conduct this investigation.

In addition, the consequences perceived by school leaders are particularly important because, even though the program theory underlying the NTES accountability policy is weak in specifying school level effects, it is at school where most of the evaluation procedures actually take place and where most of the general consequences should be perceived. The investigation distinguished between three types of consequences (negative effects, positive effects and active uses) as perceived by principals and pedagogical experts and occurring in schools at an institutional level, at teacher level, and for school leaders themselves.

The results show that school leaders (the school principal and pedagogical expert) perceive NTES as having numerous positive effects and uses at the school level and mixed effects (both positive and negative) at the teacher level. In contrast, their reports pay little attention to the uses of the assessment results by teachers and the effects regarding their own practice. Also, there are significantly more responses regarding consequences at the teacher level than at the school level, which is consistent with the individual nature of the assessment system.

At the school level what stands out are the accounts about positive uses given to the evaluation results as input for reflection, as a diagnostic of teaching quality, and as the basis for recognition. It is important to note that school leaders are not necessarily conscious of these uses as direct consequences of the NTES and that they only surface when the interviewer prompts school leaders about specific actions taken based on the information received. School leaders also report positive effects taking place at the school level such as fostering teamwork and providing external validation of the work done by the school leadership. The most important negative effect reported at the school level is the need to allocate special time for teachers to work on the NTES, which prevents them from dedicating sufficient time to their regular responsibilities.

At the teacher level, school leaders report negative effects more frequently than at the school level and these include negative emotions during the evaluation process and in reaction to poor results, as well as work overload and resistance to the evaluation process. On the positive side, school leaders perceive the NTES to promote collaboration among teachers and the revision of teaching practices as a result of teachers learning about the teaching standards. This seems to be due to their interaction while preparing the NTES instruments or due to developing increased awareness of being part of a profession that is subject to accountability.

Similar to our findings, Heneman et al. (2006) report teachers' positive responses regarding the impact of standards-based evaluation processes on their instructional practices. However, as Lustick and Sykes (2006) illustrate, these positive effects vary individually regarding their nature and potential for lasting change. Both studies attribute as the main cause of change the underlying teaching standards. The National Research Council (2008) concludes that the scarce research that has investigated the effect of NBPTS certification on teaching practices does not provide definite answers. Nevertheless, teachers who underwent the certification process tended to report that it was a worthwhile experience that improved their teaching practices and motivated them to reflect more about their practice. Similar findings come from a teacher background questionnaire that accompanies all NTES portfolios each year, which includes a few items on the perceived utility of the assessment process in the eyes of evaluated teachers (see Sun, Correa, Zapata, & Carrasco, 2011).

The explanatory analysis shows the important role of school leaders, particularly principals, for positive effects and uses to take place at the school level. This finding is mirrored in existing studies on standards-based teacher evaluation (see Halverson et al., 2004; Kimball, 2002; Koppich et al., 2006). In our case we observe accounts of more positive effects and uses in schools where the principal (i) plays a role of pedagogical leader, (ii) provides internal legitimacy to the teacher assessment process and results, and (iii) there is a positive school climate that allows teachers to share their evaluation results more openly.

In contrast to the investigation of perceived consequences at the municipal and at the individual level, we cannot make claims about whether the NTES policy is working as intended at the school level because there is no explicit description of intended consequences at this level. All we can do is to describe what school actors perceived to be actually occurring or not occurring in their schools. From what they observed, however, some important consequences expected at municipal and individual teacher levels seem to be taking place in schools as the institutions where the assessment actually happens and where assessment results are reviewed. In addition, municipal and school actors coincide regarding the NTES' unexpected consequences.

The majority of school leaders as well as a number of municipal actors declare (i) to use the NTES results as input for the provision of local rewards, (ii) that the NTES has increased collaboration among peers, (iii) that the NTES has resulted in improved teaching practices as consequence of teachers' reflection, and (iv) the use of the NTES as basis for diagnosing teacher quality. It is worth highlighting that neither many municipal actors nor many school leaders mentioned observing improved student learning as consequence of the NTES, which is the intended long-term goal of the NTES.

One important unintended but positive consequence reported by educational actors at the municipal level was that municipalities created support mechanisms for teachers to cope with the evaluation process and the effects its results generated. This is also observed – though maybe to lesser extent – at schools, since the pedagogical expert in many schools reported playing a support role – both psychologically and pedagogically – during the assessment process. School leaders and municipal actors also coincide when describing the NTES' negative (unintended) consequences: (i) increased teacher workload, (ii) teachers' negative emotions associated with the NTES such as stress and anxiety, and (iii) teachers' resistance and refusal to participate in the NTES (however, diminishing over time).

Both municipal actors and school leaders also discussed the legitimacy of the assessment and its instruments, or lack thereof, both in their own eyes and in the eyes of the evaluated teachers. They also commented on the consistency, or lack of consistency, between teachers' own impressions of their performance and what the assessment indicates, or in the case of school actors, their own impressions of their teachers' quality and the assessment results they obtained. We considered "legitimacy", as perceived by school actors, to explain why in some schools we observed more positive effects and stronger uses than in others. Furthermore, the explanatory analyses from the school and municipal levels coincide on the important role played by people holding positions of authority within the school or the local community. While at the school level the principal is key for positive effects and uses to take place, at the municipal level we observe that it is the capacity of the municipal actors to overcome the initial hostility and make effective use of the opportunities associated with the NTES. In that sense, our results highlight the importance of policy buy-in and legitimacy awarded to the assessment by the leadership of the school and local district.

Our study is limited by the fact that it only reflects school leaders' responses, which may or may not share the perspectives of the evaluated teachers. On the other hand, school leaders may be more objective observers of teacher level effects and uses than teachers themselves. In the near future we will complement the findings from municipalities and schools with information reported by teachers themselves via focus groups and personal interviews. In order to check the generalizability of our findings it would be interesting to conduct a survey about the NTES' consequences in a larger sample of school leaders. In addition, future research may include the investigation of actual consequences, as opposed to perceived consequences, by directly observing changes in school-level administrative processes and collaborative climate, coaching and preparatory activities for the evaluation, teacher evaluation culture, coordination between school and district level, as well as professional development design, implementation and impact on curriculum, instructional practices, classroom material, and student learning (for more details on a framework for evaluating the consequences of assessment programs see Lane et al., 1998).

Conclusions

This paper presents the results of our research about the perceived consequences of the Chilean national teacher evaluation system (NTES) in elementary schools. The study shows that school leaders perceive the NTES to benefit both schools and teachers in various ways. At the same time, the school leaders we interviewed reported that this teacher accountability policy also triggered strong negative reactions by teachers. These seem to have diminished over time, as teachers get more used to being held accountable and to receiving performance assessment results, and as they become more familiar with the assessment procedures and instruments.

According to school leaders, NTES results are used (i) as basis for recognition of good teachers, and (ii) as diagnostic of teacher quality. NTES also (iii) strengthens collaboration among teachers and (iv) improves teaching because the evaluation process gets teachers to reflect on their practice. Importantly, in the majority of schools such reflection is said to happen both at an individual teacher level, as well as coordinated at institutional level. Some school leaders reported using the assessment in additional ways, such as capitalizing on high-performing teachers as mentors, assigning them to difficult classes and using the results for marketing purposes.

Thus, our research so far shows a mixed but overall positive balance as to the consequences of this national teacher performance assessment system as perceived by school leaders. Both previous research regarding the municipal level as well as this study in schools show that actors from the education sector perceive that the assessment system has achieved an important number of either intended or unintended positive effects and uses, and these are especially pronounced at school level. As a next step in the research process, these findings will be contrasted with evidence collected directly from teachers.

Implications for teacher evaluation policy

We think this study provides lessons on ways that the national and local authorities could strengthen the positive perceived effects and uses of the NTES, or of any other similar large-scale standardized, high-stakes teacher evaluation system. We observed that increased use goes hand in hand with the legitimacy that the NTES has in the eyes of school leaders. *The more school leaders feel active participants of the evaluation process, the more legitimacy they might give to it and the more they might use the information.* Thus, more use would result if school leaders were involved more actively in the evaluation process, for example by increasing the weight, even moderately, of the supervisor assessment in the final NTES composite score, and by providing school leaders with more autonomy to attach consequences to the overall NTES result. Increased information and training on the NTES assessment procedures and instruments, specifically, and instructional practices and performance assessment, however, should accompany this increased responsibility. For this to happen, principals themselves may need to be evaluated based on their roles as instructional leaders instead of as administrative personnel (see Koppich et al., 2006).

For its credibility it is also important to further *strengthen the formative aspects* of the teacher evaluation system. A related program offers mandatory professional development to teachers evaluated as showing basic and unsatisfactory performance. These professional development programs need to be of high quality in order to effectively support the formative purpose of the evaluation. In addition, the evaluation should provide teachers with useful and in-depth information about their strengths and weaknesses, as well as resources to improve their pedagogical practices. Steps toward improving the individual reports have already been taken but more could be done in extending and personalizing the information the evaluation provides.

Similarly, it is important to *minimize the negative effects of the NTES*, or of any high-stakes teacher evaluation system. One important aspect concerns the excessive teacher workload and the need to provide teachers with time within the school day to prepare the NTES instruments (as well as to attend the professional development). Although some schools and municipalities already provide this time, it is not mandated and the final decision will greatly depend on municipal decision makers and principals. If teachers lack sufficient preparation time, then this may negatively affect their teaching during assessment periods, as well as their emotional experience and perceived levels of stress.

National and local authorities as well as NTES implementers should also more actively *communicate the intended and unintended effects and uses of the assessment* at the different levels of the educational system. They should also clarify what are appropriate and inappropriate preparation activities, thus contributing to the perceived legitimacy of the evaluation system within the education community.

Acknowledgements

This research was funded by the Chilean government Grant Fondecyt No. 1080135. We thank Bernardita Tornero, Emilia Aguirre and Cecilia Vidal for expert research assistance, as well as three anonymous reviewers for their insightful comments.

References

AERA, APA, & NCME, (1999). *Standards for educational and psychological testing*. Washington, DC: AERA.

Avalos, B., & Assael, J. (2007). Moving from resistance to agreement: The case of the Chilean teacher performance evaluation. *International Journal of Educational Research, 45*(4–5), 254–266doi:10.1016/j.ijer.2007.02.004.

Ball, S. (2008). *The education debate*. Bristol: The Policy Press.

Brennan, R. (2006). Introduction. In R. Brennan (Ed.), *Educational measurement* (4th ed.). Westport: Praeger Publisher.

Darling-Hammond, L., Wei, R. C., & Johnson, C. M. (2009). Teacher preparation and teacher learning: A changing policy landscape. In G. Sykes (Ed.), *The handbook of education policy research*. Washington, DC: American Educational Research Association.

Forte Fast, E. F., & Hebbler, S., with ASR-CAS Joint Study Group on Validity in Accountability Systems (2004). *A framework for examining validity in state accountability systems*. Council of Chief State School Officer.

Halverson, R., Kelley, C., & Kimball, S. (2004). Implementing teacher evaluation systems: How principals make sense of complex artifacts to shape local instructional practice. In C. Miskel (Ed.), *Theory and research in educational administration* (pp. 153–198). Charlotte, NC: Information Age Publishing.

Heneman, H., III, & Milanowski, A. (2003). Continuing assessment of teacher reactions to a standards-based teacher evaluation system. *Journal of Personnel Evaluation in Education, 17*(2), 173–195.

Heneman, H., III, Milanowski, A., Kimball, S., & Odden, A. (2006). *Standards-based teacher evaluation as a foundation for knowledge- and skill-based pay* (RB-45) Retrieved from Consortium for Policy Research in Education website: http://eric.ed.gov/ERICWebPortal/search/detailmini.jsp?_nfpb=true&_&ERICExtSearch_SearchValue_0=ED493116&ERICExtSearch_SearchType_0=no&accno=ED493116.

Herman, J., & Baker, E. (2006). Assessment policy. In R. Brennan (Ed.), *Educational measurement* (4th ed.). Westport: Praeger Publisher.

Hsieh, H., & Shannon, S. (2005). Three approaches to qualitative content analysis. *Qualitative Health Research, 15*(9), 1277–1288doi:10.1177/1049732305276687.

Kane, M. (2002). Validating high-stakes testing programs. *Educational Measurement: Issues and Practice, 21*(1), 31–41.

Kane, M. (2006). Validation. In R. Brennan (Ed.), *Educational measurement* (4th ed.). Westport: Praeger Publisher.

Kimball, S. (2002). Analysis of feedback, enabling conditions and fairness perceptions of teachers in three school districts with new standards-based evaluation systems. *Journal of Personnel Evaluation in Education, 16*(4), 241–268.

Koppich, J. E., Humphrey, D. C., & Hough, H. J. (2006). Making use of what teachers know and can do: Policy, practice, and national board certification. *Education Policy Analysis Archives, 15*(7), 1–30.

Koretz, D., & Hamilton, L. (2006). Testing for accountability in K-12. In R. Brennan (Ed.), *Educational measurement* (4th ed.). Westport: Praeger Publisher.

Lane, S., Parke, C. S., & Stone, C. A. (1998). A framework for evaluating the consequences of assessment programs. *Educational Measurement: Issues and Practice, 17*(2), 24–28.

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

Linn, R. L. (1997). Evaluating the validity of assessments: The consequences of use. *Educational Measurement: Issues and Practice, 16*(2), 14–16.

Linn, R. L. (2009). The concept of validity in the context of NCLB. In R. Lissitz (Ed.), *The concept of validity. Revisions, new directions and applications* (pp. 195–212). Charlotte, NC: Information Age Publishing.

Linn, R. L., & Baker, E. L. (2002). *Validity issues for accountability systems (No. 585)*. Center for Research on Evaluation, Standards, and Student Testing (CRESST), Graduate School of Education & Information Studies, University of California Los Angeles.

Lustick, D., & Sykes, G. (2006). National board certification as professional development: What are teachers learning? *Education Policy Analysis Archives, 14*(5), 1–46.

Manzi, J., Gonzalez, R., & Sun, Y. (Eds.). (2011). *La Evaluación Docente en Chile*. The Chilean national teacher evaluation systemSantiago, Chile: Pontificia Universidad Católica de Chile.

Miles, M., & Huberman, A. (1994). *Qualitative data analysis: An expanded sourcebook* (2nd ed.). London: Sage.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23*(2), 13–23.

Messick, S. (1998). Test validity: A matter of consequence. *Social Indicators Research, 45*, 35–44.

Ministry of Education. (2004). *Marco Para la Buena Enseñanza*Framework for good teaching]. Retrieved from http://www.educarchile.cl/Userfiles/P0001/File/Marco%20para%20buena%20Enseñanza%202004.pdf.

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

Ministry of Education. (2010). *Departamento de Estudios y Desarrollo de la División de Planificación y Presupuesto del Ministerio de Educación de Chile. Sistema de Información de Estadísticas Educativas SIEE*Information system about educational statistics]. Retrieved December 24, 2010, from http://w3app.mineduc.cl/Sire/index.

Ministry of Education Legal Department (2004). *Reglamento sobre Evaluación Docente*. Santiago de Chile.

Moss, P. A. (1998). The role of consequences in validity theory. *Educational Measurement: Issues and Practice, N17*(2), 6–12.

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. In M. Hakel, J. K. Anderson, & S. Elliot (Eds.), *Board on testing and assessment, center for education, division of behavioral and social sciences and education*. Washington, DC: The National Academies Press.

NCME Newsletter 18(1).

Patton, M. Q. (1997). *Utilization-focused evaluation: The new century text* (3rd ed.). Thousand Oaks, CA: Sage.

Patton, M. Q. (2002). *Qualitative research and evaluation methods* (3rd ed.). Thousand Oaks, CA: Sage.

Peterson, K. (2000). *Teacher evaluation: A comprehensive guide to new directions and practices* (2nd ed.). London: Sage.

Popham, W. (1997). Consequential validity: Right concern, wrong concept. *Educational Measurement: Issues and Practice, 16*(2), 9–12.

Porter, A. C., Youngs, P., & Odden, A. (2001). Advances in teacher assessments and their uses. In V. Richardson (Ed.), *Handbook of research on teaching* (4th ed.). Washington, DC: American Educational Research Association.

Ravitch, D. (2010). *Death and life of the great American school system: How testing and choice are undermining education*. Philadelphia: Perseus Books Group.

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

Sato, M., Wei, R. C., & Darling-Hammond, L. (2008). Improving teachers' assessment practices through professional development: The case of National Board Certification. *American Educational Research Journal, 45*(3), 669–700.

Setliff, B. (1989). *The effects of the texas teacher appraisal system on the climate of six small school systems* (Unpublished doctoral dissertation). Texas Tech University, Education Department. Retrieved from: http://etd.lib.ttu.edu/theses/available/etd-02262009-31295005847982/unrestricted/31295005847982.pdf.

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

Sireci, S. (2009). Packing and unpacking sources of validity evidence: History repeats itself again. In R. Lissitz (Ed.), *The concept of validity. Revisions, new directions and applications*. Charlotte, NC: Information Age Publishing.

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

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.

Weiss, C. H. (1973). The politics of impact measurement. *Policy Studies Journal, 1*, 179–183.