Teacher Education, Learning Innovation and Accountability
Series Editor: Claire Wyatt-Smith

Jorge Manzi
Yulan Sun
María Rosa García *Editors*

Teacher Evaluation Around the World

Experiences, Dilemmas and Future Challenges

Springer

Teacher Education, Learning Innovation and Accountability

Series Editor

Claire Wyatt-Smith, Institute for Learning Sciences and Teacher Education, Australian Catholic University, Brisbane, QLD, Australia

This book series offers research-informed discussion and analysis of teacher preparation, certification and continuing professional learning and the related practice and policy drivers for change and reform. The series fosters and disseminates research about teaching as a profession of choice while offering a unique link to the realities of pre-service experience in workforce preparation. It takes account of research on teacher formation that opens up issues not routinely connected: what teachers need to know and be able to do, and who they are, namely the person of the teacher and their capabilities in contributing to students' personal development and wellbeing. This goal provides a current, practical and international view of the future of initial teacher education programs.

Jorge Manzi · Yulan Sun · María Rosa García
Editors

Teacher Evaluation Around the World

Experiences, Dilemmas and Future Challenges

Springer

Editors
Jorge Manzi
UC MIDE Measurement Center
Pontifical Catholic University of Chile
Santiago, Chile

María Rosa García
UC MIDE Measurement Center
Pontifical Catholic University of Chile
Santiago, Chile

Yulan Sun
Center for Professional Teacher Development
Diego Portales University
Santiago, Chile

ISSN 2524-5562 ISSN 2524-5570 (electronic)
Teacher Education, Learning Innovation and Accountability
ISBN 978-3-031-13638-2 ISBN 978-3-031-13639-9 (eBook)
https://doi.org/10.1007/978-3-031-13639-9

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Contents

Part I Challenges in Teacher Evaluations

Part II Teacher Evaluation Systems Around the World: North America

Part III Teacher Evaluation Systems Around the World: Latin America

Part IV Teacher Evaluation Systems Around the World: Europe

Part V Teacher Evaluation Systems Around the World: Asia and Oceania

Chapter 1
Introduction

Jorge Manzi, María Rosa García, and Yulan Sun

Teacher evaluation has been a highly debated issue in most countries and educational systems, where it has been implemented, discussed, or even promoted. Arising from the consistent evidence that teachers and teaching are the most relevant factor for educational improvement (e.g., Barber & Mourshed, 2007; Brophy & Goode, 1986; Darling-Hammond, 2000; Hattie, 2009; Rivkin et al., 2005; Tucker & Stronge, 2005), all over the world, countries started to review their educational policies in the last two decades, aiming at improving teaching and teacher professional development. Teacher evaluation appeared as a key component in this trend (Goldhaber & Anthony, 2007; Hallinger et al., 2014; OCDE, 2005, 2013a, 2013b). At the system level, teacher evaluation could be used to diagnose the state of teaching and identify areas of improvement and intervention, while at the individual level, the evaluation could create the conditions for professional development decisions that could promote sustained improvement of teacher's performance in the classroom. In addition, teacher assessments could be used to make decisions related to the promotion, salaries, incentives, and contractual status of teachers, replacing the traditional approach based on seniority or certification of courses.

The editors of this book have been associated to the implementation of a national teacher evaluation system in Chile that has functioned since 2003. During these years, we faced many challenges associated to the design and implementation of a large-scale evaluation. At the same time, we have witnessed the political underpinnings

J. Manzi (✉) · M. R. García
Pontificia Universidad Católica de Chile, Santiago, Chile
e-mail: jmanzi@uc.cl

M. R. García
e-mail: rosagarcia@uc.cl

Y. Sun
Universidad Diego Portales, Santiago, Chile
e-mail: yulan.sun@udp.cl

J. Manzi et al. (eds.), *Teacher Evaluation Around the World*,
Teacher Education, Learning Innovation and Accountability,
https://doi.org/10.1007/978-3-031-13639-9_1

of these evaluations and the complexities in linking the evaluation with professional development. This experience, as well our familiarity with other teacher evaluation experiences in the world, were our initial motivation for this book. To some extent, this is a book that we wished we had read at the time we were invited to provide the technical assistance for the implementation of our national evaluation.

In this book, we wanted to offer an opportunity to explore the conditions and advances that several countries and systems experience in their effort to develop teacher evaluations. We invited cases from different parts of the world: North America, Asia, Europe, Oceania, and Latin America and also, cases that were in different stages of development, including some that were in the design phase (German state of Baden-Württemberg), others involved in the implementation, and even one case in which the evaluation was terminated after a few years of implementation (Mexico). The authors of these chapters were invited to consider topics such as the purpose of the evaluation system and their theory of action, the origin of it, the framework or standards in which is based, the characteristics, instruments used, results and consequences of the evaluation, and some discussion about the lessons learned and future challenges.

At the same time, we invited contributions to discuss four challenges that are involved in most teacher assessments: (1) the formative versus summative goals of the evaluation, including the possibility of combining both; (2) the focus on teacher performance versus teacher effectiveness; (3) the political conditioning of teacher evaluation policies; and (4) the relationship between teacher evaluation and teacher professionalism and professional development. We were extremely fortunate to have been able to gather an exceptional team of scholars and policy experts, for the conceptual and system case chapters of the book. In these pages, we explain our decision to select the conceptual dilemmas presented in the four initial chapters.

Anyone familiar with educational assessments knows about the fundamental distinction between formative and summative purposes of those assessments. While this distinction was originated in the area of program evaluation, it has been widely adopted in the area of assessments and evaluation. The chapter by Bell and Kane offers a rich contextualization for this distinction, using the framework of Weick (1976), which differentiates loosely coupled and tightly coupled educational systems. At one end of the continuum, loosely coupled schooling and teacher preparation systems have decentralized control over the goals of schooling. Multiple organizations within the nested schooling system exert control over what is taught and how it is taught. This end of the continuum is characterized by strong norms of individual teacher accountability for teaching processes and student outcomes. It is also defined by relatively weaker norms and less agreement about what counts as "good teaching". At the other end of the continuum, there is more centralized control over the goals of schooling. Contexts are defined by having higher levels of collective accountability for teaching processes and outcomes. Country contexts at this end of the continuum are also characterized by stronger agreement on norms around what counts as good teaching.

Based on this conceptual framework, Bell and Kane discuss the role of formative and summative approaches to teacher evaluation. They argue that in a tightly coupled system, direct observations of teacher performance can provide a solid basis for both formative and summative evaluation. In loosely coupled systems, summative teacher evaluations will tend to be more formal and standardized, with accountability playing a major role. Both approaches to the evaluation of teaching share same ultimate goal—improving the quality and outcomes of education—but they seek to achieve this goal in different ways. Formative evaluation of teaching is designed to improve the effectiveness of individual teachers by helping them to improve their performance. Summative evaluation of teachers is designed to improve the overall effectiveness of the teachers in the system by recognizing and rewarding good teachers and by encouraging below-average teachers to improve their performance.

Having clarified that the formative of summative purposes of teacher assessments does not directly lead to specific focus of evaluation (such as addressing teacher performance versus teacher effectiveness) and does not necessarily require specific methods (such as observation of teacher versus evidence of student learning), it was relevant to address the issue of focus and method in a separate chapter, which is done by Martínez and Fernández. They analyze how teacher evaluation usually requires a taking a position around two related but distinct targets of assessment: teacher performance (what teachers do) and teaching effectiveness (the effect teachers have on specific student outcomes). Teacher performance relies on models and frameworks that outline the ideal competencies, practices, and attitudes of teachers. In contrast, teacher effectiveness focuses on measuring and improving outcomes, usually attaching incentives to the evaluation, with the expectation that those consequences will promote the improvement of instruction.

While effectiveness is often linked with summative goals and performance with formative objectives, the authors propose that the more useful distinction is at the level of specific instruments or measures, which may be more conducive to formative or summative uses. For example, classroom observation protocols tend to be used in formative teacher evaluation, since they are a source of direct evidence of teaching as it occurs in the classroom, which can be used to identify areas of improvement and professional learning. Conversely, value-added models are seen as more summative in nature, as they focus on teachers' ability to improve student outcomes and do not directly offer evidence to guide professional learning or improvement. Since most teacher evaluation systems declare an interest in both formative and summative goals, the most prevalent solution has been to rely on multiple measures.

The key conceptual issues we have mentioned this far have been prominent in most teacher assessment experiences, as reflected in the chapters that present specific systems around the world. However, it is not possible to understand the nature of those systems considering only those conceptual definition. Political actors and their power are essential to understand the nature, purposes, and prospects for teacher assessments (Corrales, 1999). Zorrilla and Martínez show the extent to which educational policies and teacher assessments are shaped by political actors (with a key role of teacher unions), who have different and often changing goals (especially when a new government is appointed). Political agreements in contested educational policies are usually hard to reach and face frequent challenges and threats, as it is dramatically clear in the case of the Mexican educational reform (where teacher evaluation was the most sensitive and controversial component).

Finally, Ávalos discusses the connection between teacher evaluation and teacher profession. The chapter identifies some negative implications of teacher evaluation on teacher professionalism and professional responsibility. Especially in the context of evaluation systems with accountability purposes (for example, the use of value-added measures), Avalos warns about the erosion of professionalism that those approaches may convey. The author recommends responsible accountability (UNESCO, 2017), as an alternative basis for assessments. Within this approach, appraisal is anchored on respect for teachers as knowledge professionals. As expressed by Whitty (2000), to move in this direction requires demystifying teacher professional work. It requires teaching to be more democratic in its construction and appraisal, with teachers, parents, students, and the community as participants, thus, counterbalancing the narrow accountability demands operating in the context of market competitiveness (Whitty, 2000).

Teacher evaluations are clearly more complex, diverse, and technically challenging than traditional assessments in education, such as student achievements tests. Unlike these tests, which usually focus on a narrow set of curriculum areas (mathematics, language, and science in most cases), relying mostly on standardized instruments, teacher evaluations cover a wide range of aspects of teacher performance and effectiveness, using various assessments tools. Moreover, they usually declare formative and summative goals and are included in diverse contexts, from school or local management units to nationwide systems that in some cases are connected to salaries, incentives, and career decisions. Moreover, those who are evaluated represent a social and political force, which is not usually the case in other educational evaluations such as student assessments. These are just some of the differences that could be found in countries or systems that have designed, developed, or implemented these evaluations. The cases included in this book exemplify the variability of teacher evaluations around the world, demonstrating the many cultural, political, and educational underpinnings of those experiences. They also show that the development of validation studies is a pending challenge in most evaluation systems. In fact, even do we ask authors to refer to validation studies, most chapters do not provide evidence about validity of those systems. Having external agencies that can review the teacher evaluations, collect validity evidence from different sources, have resources allocated for it, and thus contribute to the well-founded improvement of them is still a debt.

The conceptual chapters discuss some of the key issues and challenges that make it possible to understand the characteristics of specific systems in different parts of the world: Europe, North America, Latin America, Asia, and Oceania. We hope the chapters in this book will help researchers, practitioners, and decision-makers in the field, especially those involved in the design, review, implementation, and validation of teacher evaluations.

1.1 In Memoriam

Before concluding this introduction, we would like to express a few words in memory and gratitude to Margarita Zorrilla, outstanding Mexican educator and researcher, co-author of Chap. 4 of this book, who passed away last January 20. With the generosity and courage that were characteristic of her, in 2020, Margarita accepted our invitation to write this chapter, while facing a serious illness, and perhaps because of the anticipation of what could happen, she proposed to do it "in four hands", with her colleague and friend Arcelia Martínez.

Writing a book takes time and in the almost two years it took to complete this book, we lost Margarita. As colleagues and friends, we were privileged to know her passionate commitment to the improvement of education and the vigorous will with which she dedicated her professional life to that endeavor. Whether as a researcher, as a teacher educator, holding prominent positions in the Mexican educational system, or in a dialog with classroom teachers, Margarita always showed her unusual combination of critical attitude and unwavering enthusiasm, academic rigor, and political sensitivity. She was an energetic fighter and a friend of infinite warmth.

Her contribution in this book only adds to the extensive legacy that she left for education in her country and in Latin America.

References

Barber, M., & Mourshed, M. (2007). *How the world's best-performing school systems come out on top*. McKinsey & Company.

Brophy, J., & Goode, T. L. (1986). Teacher behavior and student achievement. In M. C. Wittrock (Ed.), *Handbook of research on teaching* (3rd ed., pp. 328–375). MacMillan.

Corrales, J. (1999). *Aspectos políticos en la implementación de reformas educativas*. Santiago de Chile: PREAL.

Darling-Hammond, L. (2000). Teacher quality and student achievement: A review of state policy evidence. *Education Policy Analysis Archives, 8*(1), 1–44.

Goldhaber, D., & Anthony, E. (2007). Can teacher quality be effectively assessed? National board certification as a signal of effective teaching. *The Review of Economics and Statistics, 89*(1), 134–150.

Hallinger, P., Heck, R. H., & Murphy, J. (2014). Teacher evaluation and school improvement: An analysis of the evidence. *Educational Assessment, Evaluation and Accountability, 26*(1), 5–28.

Hattie, J. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. Routledge.

OCDE. (2005). *Teachers matter: Attracting, developing, and retaining effective teachers*. OCDE Publishing. Retrieved from https://www.oecd.org/education/school/34990905.pdf

OCDE. (2013a). *Synergies for better learning: An international perspective on evaluation and assessment*. OCDE Publishing. Retrieved from https://doi.org/10.1787/9789264190658-en

OCDE. (2013b). *Teachers for the 21st century: Using evaluation to improve teaching*. OECD Publishing. Retrieved from https://www.oecd.org/site/eduistp13/TS2013b%20Background%20Report.pdf

Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, schools, and academic achievement. *Econometrica, 73*(2), 417–458.

Tucker, P. D., & Stronge, J. H. (2005). *Linking teacher evaluation and student learning.* Association for Supervision and Curriculum Development.

UNESCO. (2017). *Global monitoring report 2017/8: Accountability in education: Meeting our commitments*. UNESCO.

Weick, K. (1976). Educational organizations as loosely-coupled systems. *Administrative Science Quarterly, 21*(1), 1–19. https://doi.org/10.2307/2391875

Whitty, G. (2000). Teacher professionalism in new times. *Journal of in-Service Education, 26*(2), 281–295. https://doi.org/10.1080/13674580000200121

Part I
Challenges in Teacher Evaluations

Chapter 2
Formative and Summative Teacher Evaluation in Social Context

Courtney Bell and Michael Kane

Abstract Formative and summative teacher evaluation systems are tightly connected to their country contexts. This chapter offers a framework for understanding and building formative and summative teacher evaluation systems. Building on Weick's (Weick, Administrative Science Quarterly 21:1–19, 1976) description of loosely and tightly coupled systems, we identify four contextual aspects of educational systems that play an important role in the structure and function of teacher evaluation systems. These four aspects are (1) the goals of teaching, (2) the shared understanding of good teaching, (3) the degree of centralized control over curricula and teaching practices, and (4) the structure and norms around improving teaching. The chapter also explains how a theory of use might help to clarify formative teacher evaluation systems and offers criteria by which summative teacher evaluation systems might be judged. Ultimately, the question of how formative and summative teacher evaluation should be structured is complex. The country context and its embedded systems will determine how evaluation systems might best be constructed and used for the improvement and monitoring of teaching. The chapter's treatment of teacher evaluation illuminates many of the issues researchers and practitioners might productively consider as they seek to understand and improve formative and summative teacher evaluation systems.

2.1 Introduction

In international work on teaching, it is tempting to assume that assessment systems and insights that work well in one country can "travel" to another country. If we decide, for example, that derived measures of student learning (e.g., value-added

C. Bell (✉)
University of Wisconsin, Madison, WI, USA
e-mail: courtney.bell@wisc.edu

M. Kane
ETS, Princeton, NJ, USA
e-mail: mkane@ets.org

J. Manzi et al. (eds.), *Teacher Evaluation Around the World*,
Teacher Education, Learning Innovation and Accountability,
https://doi.org/10.1007/978-3-031-13639-9_2

measures) can make a useful contribution to the summative evaluation of teaching in the United States, we may be tempted to assume that summative evaluations of teaching in other countries will benefit from the inclusion of these derived measures. After all, surely everyone cares whether teachers are improving students' learning outcomes. However, such assumptions do not adequately account for the reality that schooling outcomes are the product of social systems, which vary in the value assigned to different goals for learning (e.g., cognitive vs. social/emotional, basic skills vs. conceptual development). In contrast to biological or chemical systems, social systems operate on principles connected to countries' social and historical contexts. And it is these social–historical contexts that shape how any teacher assessment system will function to achieve valued learning goals.

Our chapter rests on the recognition that countries differ from one another in social–historical context. Further, the connections between and among K-12 schooling systems, government regulatory systems, and higher education systems where teachers are prepared also vary. Thus, both social–historical context and education system connections must be considered when specifying sensible formative and summative teacher evaluations.

The chapter offers a framework for understanding and building teacher evaluation systems. We begin by explaining how country contexts vary along a continuum from loosely coupled systems to more tightly coupled systems (Weick, 1976). Contexts characterized by loosely coupled systems tend to be decentralized and variable in how schools and teaching are organized and evaluated and have administrative authority distributed across system levels and people, while contexts with tightly coupled educational systems have more centralized control over curricula and teaching practices, but ironically, these systems may allow more local control over teaching practices, evaluations of teaching, and staff development (Weick, 1976). Educational contexts differ across countries and may also differ across regions within countries, and we explore four aspects of these contexts that are particularly critical to successful teacher evaluation systems. One important, though somewhat understudied aspect of teacher evaluation systems is the distinction between good and effective teaching (Fenstermacher & Richardson, 2005). Good teaching is defined in terms of practices that are known to be generally effective, while effective teaching is defined in terms of the outcomes of specific instructional interventions. This distinction implicitly and explicitly shapes evaluation systems (e.g., the observational rubrics, the questionnaires used to understand students' perspectives on teaching quality) in ways that reflect the social context in which the evaluation system operates.

Paying attention to system connections and conceptions of teaching quality, we specify differences between formative and summative evaluation systems. *Formative evaluations of teaching* are intended to help teachers improve their teaching by providing them with feedback on their performance and suggestions on how to improve their performance. *Summative evaluations of teaching* are intended to evaluate the teacher's performance with the goal of making decisions about the teacher. Using these definitions, we describe the criteria against which formative and summative systems might be evaluated. This treatment of teacher evaluation can be used to illuminate many of the issues researchers and practitioners might productively consider as they seek to understand and improve formative and summative teacher evaluation systems.

2.2 Social and Organizational Contexts

Chile's organization of schools and the teaching profession is not the same as England's, nor is it the same as Japan's. Each country's context serves as the backdrop for the teacher evaluation policies that can be imagined and implemented. As a case in point, the United States' systems include grade-level end-of-year student assessments and strong individual teacher accountability for students' academic learning. In the U.S. context, the use of value-added measures (VAM) as a part of yearly summative teacher evaluation appears logical to many stakeholders. But the use of VAM scores as a summative teacher evaluation measure would be less logical in the Finnish context, which does not have a system of end-of-year student tests and has a strong norm of collective responsibility for the teaching profession (Sahlberg, 2007). Trans-national effects associated with the work of various international organizations (e.g., OECD) tend to encourage increasing similarity in educational systems around the world, but elementary and secondary education are still very responsive to local and national norms and traditions. In short, context matters.[1]

In this section, we describe a continuum of social and organizational contexts for teacher evaluation. We treat context at the country level, considering systems within that context. As citizens of the United States, we are aware that Massachusetts and Mississippi have very different systems and outcomes for K-12 schooling, despite being in the same country (c.f., Peterson et al., 2011). Clearly, a country-level treatment of context has its limitations. However, to make the chapter manageable, we work at the country level and acknowledge the argument might be better applied at the level of a local geographic unit (e.g., a state or large school district in the United States or rural and urban schools in Chile, or a country in the case of Hong Kong). For a discussion of how teaching writ large varies by country and other units of aggregation, see Paine et al. (2016).

Teacher evaluation contexts can be characterized along a continuum. Across the continuum, there are social and historical forces that shape the ways people interact with one another. Within those contexts, educational systems are connected to one another more or less tightly. On one end, schooling and teacher preparation systems can be loosely coupled (Weick, 1976). Loose coupling suggests that events within the system are "responsive, but that each unit within the system also preserves its own identity and some evidence of its physical or logical separateness" (p. 3). Contexts on this end of the continuum are also characterized by having decentralized control

[1] There are also transnational influences on teacher evaluation and assessment more broadly. For example, organizations that seek to influence higher education and primary and secondary schools at the national level (e.g., OECD, the World Bank, UNESCO, etc.), research organizations that run multinational comparative studies (e.g., IEA, RAND), and influencers that create reports (e.g., McKinsey & Co, the Gates Foundation) also shape teacher evaluation systems around the globe.

over the goals of schooling. When there is decentralized control over goals, multiple organizations within the nested schooling system exert control over what is taught and how it is taught. This end of the continuum is characterized by strong norms of individual teacher accountability for teaching processes and student outcomes. It is also defined by relatively weaker norms and less agreement about what counts as "good teaching." At the other end of the continuum, teaching practices in tightly coupled evaluation contexts tend to be more responsive to one another. In these contexts, there is more centralized control over the goals of schooling. For example, teachers may all use the national curriculum and its associated textbooks. Contexts are defined by having higher levels of collective accountability (e.g., a school, a grade-level team) for teaching processes and outcomes. Country contexts at this end of the continuum also are characterized by stronger agreement on and norms around what counts as good teaching (Fig. 2.1).

In this section, we elaborate this continuum and note the connections among four key aspects of countries' contexts that are especially relevant for understanding formative and summative teacher evaluation. Those aspects are the goals of teaching, shared understanding of "good teaching," degree of centralized control over curricula and teaching practices, and the structure and norms around the improvement of teaching. After explaining each aspect, we use the cases of Singapore (tightly coupled) and the United States (loosely coupled) as examples of contexts at the endpoints of the continuum. Many—perhaps most—countries are best located between the two endpoints.

Ironically perhaps, the more general agreement on the goals of schooling and teaching practices and the more centralized curricula in tightly coupled educational systems allow for a more decentralized, local approach to the evaluation and improvement of teaching. In loosely coupled systems, it is not possible to use evaluation systems based on shared norms and shared conceptions of good teaching, because

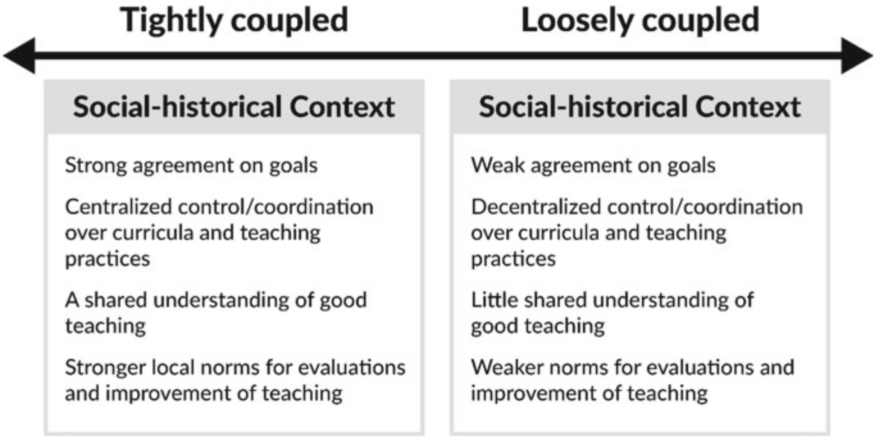

Tightly coupled
Loosely coupled
Social-historical Context
Strong agreement on goals
Centralized control/coordination over curricula and teaching practices
A shared understanding of good teaching
Stronger local norms for evaluations and improvement of teaching
Social-historical Context
Weak agreement on goals
Decentralized control/coordination over curricula and teaching practices
Little shared understanding of good teaching
Weaker norms for evaluations and improvement of teaching

Fig. 2.1 Continuum of teacher evaluation contexts

there are, at best, weak norms and limited agreement on what constitutes good teaching.

2.2.1 Goals of Teaching

All schools have goals. These goals are value judgements about what is most important for the students and society. David Labaree (1997) articulated three schooling goals that are held in tension in the U.S. capitalist democratic society: democratic equality, social efficiency, and social mobility. The democratic equality goal of schools suggests that schools should function to teach students the knowledge, capabilities, and beliefs that are needed to participate in the democracy as an informed citizen. This is important in part, because a democracy depends on an educated voting citizenry for sound governance. Schools must also teach students the skills needed to participate effectively in the economy. This means schools should prepare children for varied roles—as nurses, accountants, electricians, etc.—that reflect the society's need for an efficient, differentiated economy. Finally, U.S. schools must provide each individual student with the means of social advancement. Broadly, this goal is met when schools help each student achieve the educational outcomes (e.g., knowledge, capabilities, credentials, and access to resources) that allow for success in the social system. The first two goals tend to focus on public aspects of educational goods, while the last goal focuses on the individual benefits that accrue to students from their schooling.

Over years and across communities, these goals are in tension with one another. At specific moments in the U.S. democracy for example, policy makers created policies that supported new immigrants to prepare to become democratic citizens, while at other moments such as now, policies are more focused on supporting all students to advance socially. Labaree articulated these goals in the U.S. context, but his insights illustrate at least two points about schooling goals around the world. There are multiple, possibly competing goals for public schools in any society. Given the diversity of societies around the world, we should expect that teacher evaluation systems will be aligned to schooling goals that vary by country and over time.

In every society, the goals of schools serve as the broad, abstract targets that teachers aim for in their interactions with students and that get translated into the schools' policies and processes. They frame what is taught (the curriculum) and who is taught that curriculum at various levels of mastery and depth. For example, in a society in which all students must be able to vote for candidates who will govern the nation regarding issues such as global warming policies, schools might require science learning for all students from kindergarten to high school. If this were true, formal teacher learning opportunities[2] and professional teaching standards might in turn focus on teaching practices that require students to work independently and

[2] We refer here to teacher learning across one's professional career. Teacher education programs, i.e., the professional pre-service programs that prepare teachers for careers, play a critical role in the

in groups to analyze scientific information and make accurate inferences regarding environmental policies. Regardless of the specific schooling goals valued in a specific country, those goals will inform the vision of “good teaching” in the community.

The society’s schooling goals also specify the degree to which similar outcomes are expected for all students. Such goals of equitable student outcomes in turn shape the teacher evaluation system. For example, in the United States, there is a pervasive national belief that all children should be given the opportunity to learn and succeed academically. This belief is built into teaching and learning policies such as regulations requiring that students with special learning needs (e.g., those with autism spectrum disorder, those with dyslexia, or attention deficit disorder) have individual educational plans and be given access to the least restrictive environment that allows them to achieve their learning goals. This goal—and the value placed on it—is visible in the teacher evaluation system.

In the U.S. state of Georgia, with its diverse population that includes students with special needs, those from historically disadvantaged racial and economic backgrounds, and multilingual learners, teachers are expected to “close the gaps” between students with these backgrounds and students from White, middle-class backgrounds who speak English as their first language. The school accountability system includes two measures for this goal at the school level, the Beating the Odds metric that compares schools to other schools serving similar children (The Governor’s Office of Student Achievement, 2017), and a “closing the gap” metric that measures a school’s year-over-year progress in helping students from these groups achieve increasing levels of academic proficiency (Georgia Department of Education, 2018). These school-level goals in turn shape teacher evaluation metrics that measure, for example, whether an individual teacher works with special education colleagues to address students’ varied needs, and the quality of a teacher’s ability to set high expectations for all children and differentiate instruction appropriately given students’ background knowledge.

2.2.2 Centralization of the Education System

The degree to which curricula, definitions of good teaching, and teacher learning are centralized within a country sets the policy context for the nature and goals of the teacher evaluation system. Countries such as Australia and the United States have relatively decentralized schooling systems; local districts may choose the curriculum from which teachers teach as well as the associated teaching practices. In the United States, the decentralized K-12 system is accompanied by a decentralized teacher learning system in which teacher preparation programs vary widely in both structure and curricular focus, from for-profit alternative certification programs to university-based four-year undergraduate programs to district-led teacher residency programs

context of formative and summative teacher evaluation. An independent treatment of those complex preparation institutions is beyond the scope of our chapter.

(Cochran-Smith et al., 2016). Teachers prepared in these diverse programs are required to meet basic knowledge requirements (Gitomer & Zisk, 2015), increasingly common performance requirements (c.f., Bell et al., 2018; Darling-Hammond & Hyler, 2013), and other state requirements (e.g., a semester of student teaching, a course on CPR). However, variable programs combined with curricular differences produce teachers that do not begin their careers as teachers with the same knowledge and skills (Boyd et al., 2009), and can then be associated with different learning rates based in part, on the school environment (Papay & Lasky, 2020; Sass et al., 2012). That is, in the United States, schools and teacher preparation programs are both loosely coupled systems, and these two systems are only loosely coordinated.

Countries that have more centralized control over schooling and teacher learning may provide little opportunity for local districts to select the curriculum or teaching practices used in classrooms. Centralized control of what students and teachers learn may be accompanied by similar levels of teaching quality across classrooms and little difference in teaching quality between novice and veteran teachers. The recent TALIS Video study showed exactly these patterns in Shanghai, a more centralized system in which curriculum and teacher learning have little local variability. That study found that in the quality of classroom management practices, seven participating countries had variability in average observation scores of roughly 0.19 points on a four-point scale. Shanghai classroom to classroom variation was less than a third of this variation—just 0.06 points (Bell et al., 2020). Further, the study found that even after controlling for student background characteristics, there was no relationship between the quality of classroom management or instructional practices and teachers' years of experience in Shanghai (Doan & Mihaly, 2021). Countries with more centralized systems of schooling and teacher preparation may have more homogenous policy contexts in which teacher evaluation policies are designed and implemented.

2.2.3 Shared Understanding of Good Teaching

Given their agreement on goals, curricula, and general patterns of instruction, tightly coupled systems are likely to develop shared understandings of teaching practices that they consider to be likely to achieve the goals within the framework of the common curricula, and general patterns of instruction. If all teachers are pursuing essentially the same goals, within the context of common curricula, and in the same kind of instructional context, they may achieve a fairly high level of general agreement on how to teach. There will be variations in approaches, with some teachers spending relatively more time explaining topics to the class, and other teachers spending more time leading class discussions or working with small groups of students, but the teachers are likely to use the same set of methods in more-or-less the same ways. As a result, they can share a conception of high-quality teaching and can recognize it when they see it.

In loosely coupled systems, with little coordination of goals, curricula, and general patterns of instruction across the system, it is much more difficult to achieve widespread agreement on what good teaching looks like. If teachers are pursuing somewhat different goals, using different materials and curricula, in different instructional contexts, it will be more difficult for them to achieve a high level of agreement on how to teach. Given the variability inherent in loosely coupled systems, there is likely to be more variation in teaching.

As a case in point, in his seminal book, *Schoolteacher*, Lortie (1975) explains that teaching in the United States—a loosely coupled system—is defined by the lack of a "shared technical culture." Teachers do not agree on the teaching practices that enable successful student learning, and teacher learning is achieved through trial and error, rather than as "the refinement and application of generally valid principles of instruction" (Lortie, 1975, p. 80).

Finally, the degree to which a country's shared technical culture focuses on more abstract versus detailed teaching practices will also shape the teacher evaluation system's role. For example, a shared culture around more abstract teaching practices might include agreement around ideas and general teaching approaches (e.g., good teaching should support students' social-emotional growth) whereas a shared culture around more detailed teaching practices might focus on the application of those ideas and approaches to specific groups of children, assignments, and/or subject matters (e.g., good teaching allows 14-year-olds to self-select topics for their yearlong history projects). Formative and summative evaluation systems will vary in their function depending on the nature and specificity of a country's shared technical culture for teaching.

2.2.4 Norms for the Improvement of Teaching

Closely connected to system centralization is the infrastructure and norms around teacher learning and development. In centralized systems, teachers enact teaching practices that are similar to one another or "core" to teaching (Ball & Forzani, 2011). Almost every teacher carries out conversations with parents/guardians. All teachers explain content to students—whether standing next to a student at their desk or standing at the front of the room for all students to hear. These types of practices are general and common across teachers, and all teachers must be able to carry them out competently. Yet, teaching practices must also be sensitive to local contexts—the subjects, grade levels, and the students that teachers interact with daily. Teaching 5-year-olds is not the same as teaching 15-year-olds. Every country must then have ways to support teacher learning and development that is both general and specific.

The approach to teacher learning has both structural and normative aspects. For example, in some contexts, schools set aside times for common planning and learning for all teachers. This time may be used for collective conversations regarding improvement. Some schools have one day a month in which students are sent home for an afternoon and all teachers attend professional learning workshops. In the

United States, teachers are frequently required to engage in professional learning—e.g., taking university-based courses—in order to move higher on the salary schedule. However, what is learned may or may not be tightly connected to classroom teaching goals and performance. The resources (e.g., time, availability of courses) and their nature (e.g., individually selected, mandated to groups, aligned to evaluations or not) shape the structure of teachers' learning opportunities.

Normatively, responsibility for the quality and content of teachers' learning opportunities may reside with individual teachers and/or other groups or organizations. In Japan, mathematics teachers' learning is often organized at the school level and collectively participated in by groups of teachers focused around the study and improvement of lessons in the student curriculum (Fernandez, 2002). Responsibility for the quality of lessons is shared across the group of teachers. In the United States, teachers are provided choices in how they meet professional learning requirements and are expected to take individual responsibility for learning. Responsibility for the quality and the content of learning opportunities is driven by the market, not the government or one's peers. Teachers can choose to take a course on classroom management or on problem-based learning, as they desire. Norms about the focus, quality, and responsibility for professional learning intersect with the structures of learning opportunities to provide the context for formative and summative evaluation systems.

Singapore is a prime example of a tightly coupled educational system. Singapore has detailed student learning standards and aligned assessments that apply to all students in the country (Tee Ng, 2008). Teacher education takes place through multiple pathways at a single institution, the National Institute of Education, using a well-specified curriculum aligned to graduate teacher competencies (Goodwin, 2021). These competencies are a modification of the Ministry's evaluation and support system for practicing teachers (Darling-Hammond, 2021). Practicing teachers are guided around the National-Singapore-Teaching-Practice model (MOE, n.d.), a set of shared beliefs and pedagogical practices all teachers pledge to uphold and enact. Singapore's clarity around the goals of schooling, centralization of teacher preparation, and clear vision of good teaching is accompanied by structures and norms for teacher learning. All practicing teachers participate in a school-based professional learning community that meets weekly and engages in a group-selected learning approach—action research, lesson study, learning study, or learning circle (Darling-Hammond, 2021). Schools use replacement teachers to allow the learning communities time to meet. Teacher evaluation in this environment is implemented by school personnel yearly, is growth oriented, and has two formal goals: performance evaluation and determination of prospects for promotion in one of Singapore's three teacher career tracks (Jensen et al., 2017). Thus, in Singapore, formative and summative teacher evaluations are closely linked.

On the other end of the continuum, the United States' constitution delegates responsibility for education to the states, and within states, most educational decisions are made in local communities. As a result, the United States is a very loosely coupled system at both the national and state levels. There are no national student learning standards; however, national student goals are operationalized in one centralized student assessment (i.e., the National Assessment of Educational Progress (NAEP)) and a small number of federal mandates (e.g., that all students with special needs have individual learning plans mandated in the individuals with Disabilities Education Act (2004)). At the state level, there are multiple goals for schools, usually a single set of student learning standards assessed by a state assessment. However, this means that at the national level there are fifty different student standards–assessments combinations. Subsequently, teachers across the country aim for somewhat different curricular goals. Teachers are prepared in roughly 2000 teacher education institutions (NCTQ, n.d.), with each state having a variety of ways of regulating those institutions. There are multiple national standards of teaching practice such as the standards of the National Board for the Professional of Teaching Standards (NBPTS, 2012) and the Interstate Teacher Assessment and Support Consortium's model standards and learning progressions (CCSSO, 2013). These may or may not be used by preparation programs and many states and districts do not use a general definition of good teaching and instead, have state teaching standards. In addition, the system for professional learning is market driven, compliance oriented, and of uneven quality (Hill, 2009), with most teachers fulfilling contractually required professional development hours in teacher-selected workshops. In this context, teacher evaluation has been regulated at the state level with a heavy emphasis on summative accountability. Further, despite having the improvement of teaching as a main goal, there is little evidence that the summative and formative approaches in the United States have improved teaching. One notable exception to this is Washington, D.C. (c.f., Adnot et al., 2017; James & Wyckoff, 2020).

Singapore and the United States represent clear examples of tightly coupled and loosely coupled educational systems, but the context is more mixed in most cases with the four features of countries' educational contexts being less consistently loosely coupled or tightly coupled. It is likely then that many country contexts are combinations along the continuum. For example, Germany has similar expectations for primary school (through fourth grade) but generally has three tracks of secondary schools (Hauptschule, Realschule, and gymnasium) with different goals. Students are not all expected to achieve the same learning outcomes in secondary school because historically people believed that there were different student aptitudes that should be nurtured: manual, technical, and intellectual (Blömeke et al., 2009). These different outcomes for students are associated with different rules governing teacher training and expectations for good teaching. Interestingly, summative teacher evaluation occurs at the school level in the tradition of school inspections, which is a holistic external examination of school performance. Formative teacher evaluation is controlled at the school level and is regulated by the German states (landers) and therefore varies somewhat (Martinez et al., 2016). There will be variation in how the four contextual features of countries play out across country contexts, and this variation will shape the formative and summative teacher evaluation. More information about the United States and Germany teacher evaluation systems can be found in specific chapters of this book (see Fauth & Herbein; James, Husain & Wyckoff; Rodríguez).

It is important to note that many countries have attached consequences to both formative and summative teacher evaluation systems that further interact with the four aspects of context we note (c.f., Hallinger et al., 2014). We do not address accountability in any detail in this chapter; however, to the degree that countries attach significant consequences to teacher evaluation systems we would expect Campbell's law—the idea that when a metric becomes a policy target, it will stop being a good metric to apply, and therefore, require mitigation (Briggs, 2016).

The specific country context will also shape a country's definition of teaching quality as instantiated in its teacher evaluation system. If, for example, the goal of schooling is to support the students' ability to engage effectively in the workforce, high-quality teaching might include an emphasis on teaching skills to work individually and in groups so that students are prepared for the many ways employees must interact in the workplace. The teacher evaluation system might reasonably include ratings for the quality of flexible grouping practices or the monitoring of students.

Thus, the specification of high-quality teaching is a necessary component of both formative and summative teacher evaluation systems. This specification of teaching quality relies on a country's definition of both good teaching and effective teaching (Fenstermacher & Richardson, 2005). We now turn to this distinction and the ways in which definitions of teaching become instantiated in teacher evaluation systems through standards of teaching practice.

2.3 Good and Effective Teaching Practice

The definitions of good practice and effective practice in a profession are closely linked but conceptually distinct because they can be (and generally are) specified at different levels of generality. Good practice is conceptualized in terms of practice behaviors that are expected to be effective in most cases, and it is defined in terms of what the practitioner does (e.g., teacher behaviors). Effective practice is defined in terms of outcomes in specific cases (e.g., student learning). Good practice is defined in terms of practices that are generally effective but may not be effective in a particular case.

A particular instance of practice is said to be effective if it achieves its goals in that particular case. As an example, a class presentation would be considered good practice if it were clear, engaging, well-paced, at an appropriate level for the class, and it would be effective if students understood the material being presented. If some students did not fully understand the presentation, it would not be considered effective for those students. Note that what is effective in a particular instance of practice is highly contingent; it depends on the goal being pursued or the problem being addressed, the situation as it exists, and the context, including the instructional history of the students.

Each profession tries to identify effective ways to deal with various situations and solve various kinds of problems within its scope of practice, through research, experience, and theory. These efforts lead to the development of practices (e.g., using representations to explain mathematics concepts) that generally prove effective in dealing with certain issues that arise in practice. In many cases, extensive research is required to establish the effectiveness of specific practices for specific situations. The empirical research required to establish that a new practice is effective can take years, especially if additional research is required to determine the practice's effectiveness for different groups of students, at different grade levels, and in different contexts. How well a teaching strategy is likely to work for a particular student depends on many variables including the age, linguistic backgrounds, identities (gender, race, etc.) and interests of the student, their current level of prior achievement, and their relationship with the teacher and their classmates. The knowledge base for teaching should include this kind of information for many possible strategies, as well as a much broader domain of knowledge and strategies required for good (and generally effective) practice. For any profession, including teaching, the practice domain will generally include a very large set of strategies at various levels of generality. The practice domain will also include more detailed guidance on how to implement the strategies effectively and the circumstances in which it would be appropriate (or inappropriate) to implement them.

The *standards of practice* for a profession depend on this knowledge base as the justification for decisions about how to deal with the situations that arise in practice. The standards of practice include guidance on how to determine the nature of the problem, the general approach to take in solving the problem, and the specific actions that are known to be effective in most cases. For example, extensive research has been conducted on general components of good teaching across countries (OECD, 2020) and on effective teaching of specific subjects, like mathematics (Chazen et al., 2016), literacy (Purcell-Gates et al., 2016), and science (Windschitl et al., 2016). In addition, the practice domain will include ethical standards, legal restrictions, and cultural norms that apply at the most general level to all professional encounters.

Good practice is defined in terms of performance in ways that are consistent with the standards of practice in the profession. Good practice is expected to be effective in the sense that it usually yields good outcomes, but this is not automatic. A given instance of professional practice is said to be "good" if it is consistent with the relevant standards, but it may or may not be effective in a particular case. A surgeon may be evaluated as having performed superbly, even if the patient does not improve, or even dies. A teacher may employ carefully studied teaching strategies and do so with great skill in a particular situation with a student and not achieve the intended goal. Of course, it is expected that good practice will generally be effective, and if a particular practice is found not to achieve its goals in many cases, or has negative side effects, it may stop being used or be removed from the standards of practice for the profession.

Note that the empirical justification of the professional practices, and more generally of the standards of practice for a profession, occurs mostly at a very general level, and not at the level of particular instances of practice or at the level of particular practitioners. In the context of a particular teaching strategy, a teacher's performance is judged in terms of how consistent the performance is with the standards of practice or exhibits appropriate professional judgment (Lampert, 2001). An evaluation of the performance will focus on questions such as: Was the situation evaluated well enough? Was the approach taken appropriate, given the situation? Was the strategy implemented well? And were the outcomes followed up and any unintended outcomes dealt with appropriately? So, for example, if a teacher is making a presentation, the standards of practice would require that it be correct, clear, and well-paced. The standards would also require that the teacher attend to how the class is responding; do they seem confused or bored? Do all students understand the presentation? And of course, these standards will vary as a function of subject matter and grade level; a presentation in a high school physics class could be a lot longer and more complex than a science presentation in a second-grade class.

The *standards of practice* define good performance in the profession and serve as the basis for most evaluations of the quality of performance in the profession. For example, a physician's performance in a particular case is generally evaluated in terms of how well they implemented the standards of medical practice in the case. Their overall performance will be evaluated across all of their cases. This kind of evaluation system generally works well in medicine because there is strong agreement on a very broad set of *standards of practice* in the medical profession. The medical profession is a tightly coupled system, at least in terms of its *standards of practice*. The same is true of accounting, dentistry, law, and a number of other professions. Definitions of "good practice" get incorporated in evaluation systems for professional practice through the articulation of standards of practice that are shown (mostly through research) to generally lead to positive outcomes and that are generally accepted by the profession and the public.

Similarly, definitions of good and effective teaching get instantiated in teacher evaluation systems through the articulation of standards of practice that are known (or believed) to yield positive outcomes and that are generally accepted by the profession and the public. This tends to happen more or less automatically in tightly coupled educational systems but is difficult to achieve in loosely coupled systems, like those in the United States. And in comparison with other professions, the research base in teaching is nascent. For the conceptual distinction between "good teaching" and "effective teaching" to be implemented in teacher evaluation, it is necessary to have general agreement about the goals of schooling so that the goals that can be incorporated in standards of practice. These standards of practice do not need to be as finely articulated and extensive as those that have been developed for some other professions, but they have to be generally understood and accepted within the educational system in which the evaluation system is to be used.

Secondarily, professionals are evaluated in terms of outcomes. If a professional's practices are not successful in a large percentage of cases, this might indicate that their performance is poor and needs to be reviewed. If many of a surgeon's patients die during surgery, the quality of their work may come under suspicion, but they would not automatically be considered to be performing poorly, especially if many of their patients were considered high-risk cases. Similarly, a teacher's performance might be subjected to extra review if there were evidence that the teacher's students were not achieving the goals for the curriculum. In essentially all cases, an evaluation of the quality of a practitioner's performance will be evaluated primarily in terms of its consistency with the standards of practice. Even if poor outcomes trigger *a review* of a practitioner's work, the final evaluation will generally depend on a detailed review of their work against the standards of practice. Some teacher evaluation systems will build in reviews contingent on performance, and others might build them in as a part of the centralized control over the system.

As indicated earlier, tightly coupled educational systems tend to have high agreement on what constitutes good teaching practice, and therefore, the quality of specific instances of teaching practice can be evaluated in terms of the shared definition of good teaching specified in the *standards of practice*, with the understanding that good teaching, thus defined, will also be effective teaching. Loosely coupled systems tend to have much less agreement about what constitutes good teaching, and as a result, confidence in the link between any particular set of teaching strategies and effectiveness will be weaker. As a result, judgments about the quality of teaching may require direct evidence of student outcomes to support evaluation decisions in loosely coupled systems. Lack of agreement on quality teaching can also result in evaluation systems that are not widely trusted across levels of the system.

The presence or absence of strong agreement on a general conception of good teaching obviously has a major impact on how we design formative and summative evaluations of teaching. In particular, if a community is in general agreement on a definition of good teaching in terms of general standards of practice for good teaching, teacher evaluations are likely to emphasize classroom observations for their evaluations of teaching, and to use summaries of student outcomes on standardized tests as a secondary source of evaluative evidence. If a community does not agree on what it means to teach well, it is likely to rely heavily on summaries of student outcomes on standardized tests as a primary source of evaluative evidence and to treat direct observations of teaching as a secondary source in evaluating teaching.

Teacher evaluation systems are layered on top of conceptions of good and effective teaching. In the next section, we distinguish between formative and summative systems and discuss the connections between the two.

2.4 Framework for Understanding Formative and Summative Evaluations of Teaching

Michael Scriven (1967) introduced the distinction between formative and summative evaluation in terms of the goals of the evaluation. The original formulation was developed for program evaluation. Formative evaluations of a program would involve the collection of data during the program's development and operation in order to identify problems and to enhance effectiveness. Summative program evaluations would be conducted after the program has been developed in order to decide whether it should be implemented on an operational basis; summative evaluations may also be implemented to support decisions about the program (e.g., whether to expand the program or shut it down). Subsequently, the distinction was modified to apply to the evaluation of students (Black & William, 2003; Bloom et al., 1971; Crooks, 1988; Scriven, 1967). The distinction between formative and summative goals for evaluation is very general and can be applied in many contexts. And in particular, it can be applied to the evaluation of teaching.

As indicated, the distinction between formative and summative evaluation is one of the purposes rather than method. We can define formative and summative evaluations of teaching in terms of the intended uses of the evaluation:

> *Formative evaluations of teaching* are intended to help teachers improve their teaching by providing the teachers with feedback on their performance.
>
> *Summative evaluations of teaching* are intended to evaluate the teacher's performance over some period of teaching with the goal of making decisions about the teacher.

Both approaches to the evaluation of teaching have the same ultimate goal—improving the quality and outcomes of education—but they seek to achieve this goal in different ways. Formative evaluation of teaching is designed to improve the effectiveness of individual teachers by helping them to improve their performance. Summative evaluation of teachers is designed to improve the overall effectiveness of the teachers in the system by recognizing and rewarding good teachers and by encouraging below-average teachers to improve their performance. In extreme cases, poorly performing teachers may be removed from the classroom.

Given this difference in their purpose, the designs and implementations of formative and summative evaluations tend to differ along several dimensions, including the perceived relationship between the teacher and the evaluator, the specificity of the performances evaluated, the directness of the observations serving as a basis for the evaluation, the feedback provided to the teacher, the information provided to others, and the decisions based on the evaluation results.

2.4.1 The Perceived Relationship Between the Teacher and the Evaluator

For formative evaluations, the evaluator is likely to be a colleague, the principal, or other teaching or administrative staff, and the relationship is likely to be cooperative, with the evaluator's main goal being to provide feedback on the teacher's performance. Often formative evaluations will also provide advice on how the teacher can improve their teaching performance. For summative evaluations, the evaluator is likely to be an administrator or an external agent, charged with rendering on overall evaluation of the quality of a teacher's performance (or the performance of another unit of analysis, e.g., a school). If the summative evaluation is based on analyses of student test scores (e.g., VAMs), the teacher may not encounter an evaluator at all and instead be assigned a score by the state. In many cases, there are multiple components of the summative evaluation (c.f., Hansen & Chu, 2014), each with their own evaluators and data.

2.4.2 The Specificity of the Performances Evaluated

Formative evaluations are generally applied to specific performances that have been observed (e.g., a lesson), so that the feedback provided to the teacher can be detailed and specific to aspects of the performance that the evaluator thinks might be improved. Summative evaluations characterize some extended set of performances (e.g., over a year), so that conclusions can be drawn about the teacher's general level of performance.

This difference in the breadth of the judgments to be made has major implications for analyses of the reliability of the results of the evaluation. For formative evaluations, the results apply to a specific sample of performance and therefore do not need to generalize over classes, lessons, or the school year. For summative evaluations, the results need to be generalizable over an extended set of performances and over multiple aspects of the performance. Because the grade level or subject matter a teacher is assigned to frequently changes, summative evaluations also should attend to the stability of the evaluation results over multiple years.

2.4.3 The Directness of the Observations Serving as a Basis for the Evaluation

Formative evaluations of teaching are generally based on more-or-less direct observations of teaching performances. They may be carried out by an observer or students, through student surveys. Summative evaluations may be based on direct observations of teaching performances but may also be based on or combined with less direct indicators of the quality of teaching such as VAMs. In the previously discussed case of the U.S. state of Georgia, there are few formative evaluations of teaching; however, the summative system combines derived measures of student achievement, teacher-developed indicators of curriculum-based student learning, administrator observations of lessons, and indicators of professionalism (e.g., attendance).

2.4.4 The Feedback Provided to the Teacher

Providing helpful feedback to the teacher is the main purpose of formative evaluation. Summative evaluations generally involve little or no feedback to the teacher, and what feedback is provided is likely to come after the school year is over. Raudenbush and Rowen (2016) emphasize that the format of the feedback, numerical versus narrative, can play a large role in what a teacher understands about her performance. Both formative and summative evaluations are strengthened when they have a realistic theory of use. We turn to this matter in the next section.

2.4.5 The Information Provided to Others

Providing evaluative information about the overall quality of teaching (for individual teachers or schools) is the main focus of summative evaluations of teaching. The results of summative evaluation are generally intended to provide a basis for administrative decisions to be made by school, district, or other units with authority over the educational system. In tightly coupled educational systems with more centralization, formative evaluation may play a major role in teacher evaluation systems, and summative evaluations may play a minor role. In loosely coupled systems, summative evaluations may play a larger role and be used to exercise control over parts of the system (e.g., teachers, schools). The dissemination of information from formative evaluations is likely to be much more limited, going mainly to the teacher (and for more formal observations of teaching, to school administrators).

2.4.6 The Decisions Based on the Evaluation Results

For formative evaluations, the decisions based on the results of the evaluation tend to be local and limited to the specific teacher (e.g., that the teacher should focus on using more questioning behavior to encourage more active student participation or that the teacher should attend a continuing-education program on some topic). These decisions are likely to be made more-or-less jointly by the teacher and the evaluator. For summative evaluations, the decisions based on the results are generally more consequential for the teacher and are likely to be made by individuals (school principals, district administrators, government officials) in positions of authority. In some systems, e.g., Washington, D.C., and Georgia in the United States and Chile, summative evaluations of teachers function as accountability to the public. In these cases, severe consequences such as the revocation of a teaching license or mandatory coaching may be associated with repeated poor summative results. For summative evaluations used in such accountability systems, the evaluations need to be more formal and objective in order to survive legal and administrative scrutiny.

For tightly coupled systems with high levels of central coordination and collective responsibility for teaching processes and outcomes, formative evaluation and summative evaluation may be integrated into a common system (as is the case in Singapore). In such cases, the distinction between the two kinds of evaluation may be subtle and informal. In loosely coupled educational systems, with little central coordination and a high level of individual responsibility for teaching practice and accountability for teachers, the distinction between formative and summative evaluation may need to be explicitly defined and separately implemented, with summative evaluations playing a major role in accountability systems, as in the state of Georgia and Chile (for more information about Chile see Sun chapter in this book).

2.5 Formative Evaluations of Teaching

In formative (and summative) teacher evaluation, there should be a clear understanding of how the evaluation information will be used (King & Alkin, 2019), referred to as a "theory of use." The theory of use specifies how teacher evaluation will lead to desired changes. By specifying how evaluation is linked to change, a theory of use clarifies the goals of the evaluation and has implications for the policies and procedures necessary to implement teacher evaluation positively. A theory of use can apply at multiple levels of the system: the teacher, school, district, and country. We focus here on the teacher-level formative use.

In many countries, formative teacher evaluation is one means of strengthening student outcomes. In order for teacher evaluation to lead to changes in student outcomes, (e.g., improvements in student learning) a specific set of events would need to occur. Figure 2.2 describes one hypothetical way that information (e.g., an observer's ideas about how to improve, numerical and/or text assessments against a rubric) from a formative observation of teaching might lead to improved student learning through improved teaching practice.

Let us consider an observational measure of a formative evaluation system and begin with the first step in the theory of use in Fig. 2.2. First, the observation system—both the scales and the processes supporting the scales—must provide information (scores, narratives, observer insights) that identifies the quality of the performance at a certain level of specificity (i.e., lesson, group of students, subject matter). That level

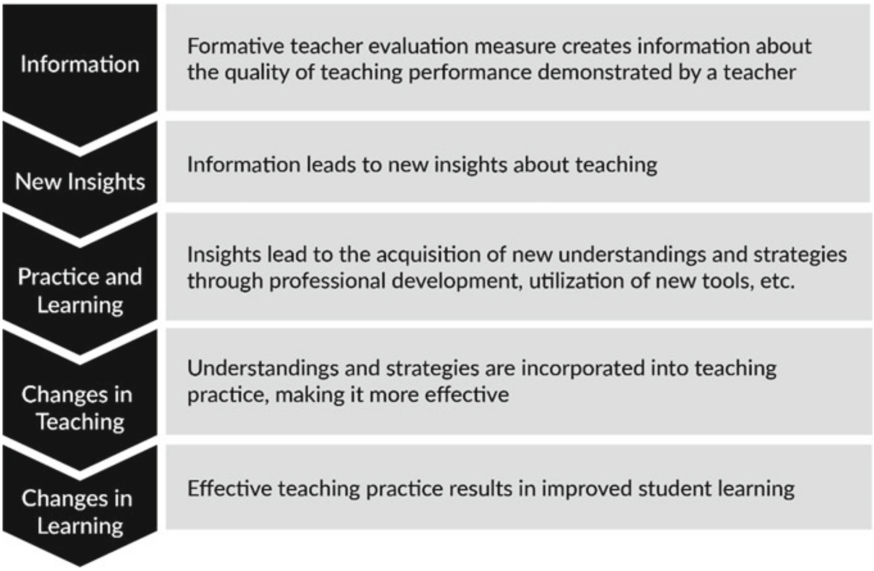

Information
Formative teacher evaluation measure creates information about the quality of teaching performance demonstrated by a teacher
New Insights
Information leads to new insights about teaching
Practice and Learning
Insights lead to the acquisition of new understandings and strategies through professional development, utilization of new tools, etc.
Changes in Teaching
Understandings and strategies are incorporated into teaching practice, making it more effective
Changes in Learning
Effective teaching practice results in improved student learning

Fig. 2.2 Theory of use linking formative teacher evaluation information to the improvement of student learning

of performance quality would be specific to the observation system. The Danielson Framework for Teaching (Danielson, 2007), for example, specifies that teaching practice should create a positive classroom environment.

The scores and narrative information, such as the notes taken by the administrator conducting the observation, would then lead to a teacher's deeper understanding of her teaching practice. For example, the teacher might realize that with respect to a positive classroom environment, she offers more negative corrective comments than positive encouraging comments when she wishes to do just the reverse. Such an insight is complex. It contains information about strengths and weaknesses, ideas about how best to improve the classroom environment, and perhaps specific instructional strategies the teacher might want to learn to use more proficiently. Insights might be developed collaboratively between teacher and evaluator. They might be individual insights or perhaps even insights that groups of teachers develop after reviewing observation scores.

Given this insight, the teacher would then take some set of actions that lead to an improved understanding and classroom strategies. That is, an effective formative evaluation system presupposes a theory of use for the results of the evaluation. Those actions might include professional learning courses or the trialing of a new planning tool in which the teacher scripts positive comments ahead of time so that she remembers to use them more frequently. Or perhaps, the teacher asks a colleague to watch, document, and give feedback on the teacher's positive and negative comments so that the teacher can learn when and to whom she offers these more negative comments. Whatever actions are taken in response to the insights about her teaching, the teacher would then incorporate what she has learned into her teaching practice, thereby making her practice more effective. To the degree practice improves in ways that are consistent with a high-quality classroom learning environment, we would expect a positive change in the learning demonstrated by that teacher's students. When such a teacher evaluation process is conducted with many teachers and students, the result would be a system-wide improvement in student learning.

There are many plausible theories of use. The one specified here is sensible for an observation system in a U.S. teacher evaluation context characterized by the goal of promoting students' social-emotional well-being, loose coupling between curriculum and teaching practices, little agreement on good teaching practices, and weak norms around improvement. This theory of use would be less sensible in a country context characterized by the goal of promoting students' academic achievement, tight coupling between curriculum and teaching standards, agreement on good teaching practices, and strong professional improvement norms. For example, in Japan or Shanghai it might make more sense for a teacher evaluation system to produce information across teachers that the head teacher(s) uses to make decisions about the focus of all teachers' professional learning in the school, thereby changing the first few steps in Fig. 2.2. Once a collective area of improvement is identified, there might be a collectively specified common approach to teacher learning and improvement, which then results in changes in teaching practice and subsequent improvements in student learning.

We have focused here on a formative theory of use and narrowly on using formative information to improve the teaching practice of an individual teacher. However, as the previous sections indicate, teachers are given feedback, information is given to others in the system, and there are consequences for both formative and summative teacher evaluation. This is true at multiple levels of the system—the school, the district, the state, and the country. This suggests that as we consider teacher evaluation systems, we consider theories of use for both formative and summative goals, as well as different system levels.

2.6 Summative Evaluations of Teaching

Summative and formative evaluations can and often do involve the same instruments but the goals are different; as such, the consequences tend to be higher for summative evaluations than they are for formative evaluations. In some systems, summative evaluations may be used to make decisions about salary, promotions, and future employment, which are subject to strict bureaucratic and legal requirements. As a result, summative evaluations of teaching generally need to be more formal and standardized than formative evaluations, especially if applied at the teacher or school level.

Countries organize their summative systems at different levels of the educational system—the school, the sector (e.g., all Catholic schools), the state, and the national level. For the remainder of the chapter, we use the example of a school-level summative teacher evaluation system because even in state and national level systems, school administrators often carry out the work of the summative evaluation. In a recent PISA 2015 report of 55 countries, administrators in the majority of students' schools carried out regular teacher appraisal. Even in some countries that have national inspectorates such as the United Kingdom, those inspectorates—Ofsted (the Office for Standards in Education, Children's Services and Skills)—issue summative school-level judgements that incorporate teacher appraisal.

2.6.1 Incorporating the Results of Formative Evaluations in Summative Evaluations

An important consideration in the design of any evaluation system is the operational level at which the evaluation is conducted, data are collected, and evaluative decisions are made. For a formative evaluation of teaching employing observations of teacher performances, the focus is typically on the teacher's performance in the classroom during a class session or part of a class session. In a tightly coupled system, with strong consensus on the definition of good teaching, the observations and evaluative suggestions made during classroom observations can play significant roles in both formative and summative evaluations. In a loosely coupled system, the formative and summative evaluation systems are likely to be distinct and possibly disconnected.

Summative evaluations of teaching are intended to provide information that will inform and support administrative decisions of some kind and are likely to be most accurately interpreted at the school level, because many of the factors that influence the quality of teaching operate at the school level. The school defines the resources available to teachers and students (the physical facilities, building leadership, the organization of time, specialized services for students with special needs, etc.). The population of students served by the school is generally determined by school or district policy (e.g., neighborhood schools, magnet schools). The teaching quality of an individual teacher or a small group of teachers will be strongly influenced by these school-level factors.

So, it is probably more useful and appropriate to focus the design of summative evaluations at the school level rather than on individual teachers, and to include the school's formative evaluation system as an integral part of each school's summative evaluation system. The quality of teaching in the school and in individual classrooms will be strongly influenced by the effectiveness of the formative evaluation in the school. The formative evaluation system can help new, inexperienced teachers to become stronger and more effective in their teaching and can help experienced teachers to maintain and improve their ongoing performance. Thus, we think that a strong system of formative evaluation that is integrated with the functioning of the school would be a basic component in any well-functioning summative evaluation system.

We recognize that the use of the results of formative evaluations as part of a teacher's summative evaluation could interfere with the effectiveness of the formative evaluation system, in that the teacher may be reluctant to reveal concerns about their performance to the evaluator, and the evaluator may feel less free to express constructive criticism to the teacher or in their report on the evaluative encounter. This problem can be addressed, at least in part, by designing the total system, and especially the formative component, to have a positive, non-punitive tone, with multiple chances to succeed, and multiple pathways to success. In any case, the gains in the effectiveness of the overall evaluation system resulting from an integrated evaluation system with a strong formative component can offset this potential consequence, especially if the system as a whole is supportive rather than punitive.

2.6.2 Five Criteria for Evaluating Summative Evaluation Systems

In this section, we will outline some of the properties that need to be given extra attention in a summative evaluation system. These properties are also needed in formative evaluations, but because of the bureaucratic purpose of summative evaluations, these properties need to be documented more thoroughly for summative evaluations.

How these properties are addressed in a particular school will depend on the social, cultural, and political context in which the school operates, and therefore, we will discuss these properties at a very general level.

Transparency: The summative evaluation system should be transparent in the sense that an external reviewer can see how the system works, and that it works.

Validity: The criteria to be used to evaluate teaching (or teachers) should be based on a definition of good teaching that is derived from research on effective teaching practices and that is shared by the teachers, the community, and school administrators. The criteria should be specified in advance and communicated to the teachers in advance and should be implemented consistently.

Fairness: The summative evaluation system should be designed and implemented in a way that is free of any identifiable sources of bias. It is generally not possible to ensure that any evaluation is completely fair, but it is possible to identify characteristics of the system that might lead to bias in the evaluation system, and to eliminate or at least ameliorate them.

Reliability/Consistency: The results of a summative evaluation system need to be reliable, consistent, and robust in the sense that they reflect the target variable, the quality of teaching. They also should not be too strongly influenced by incidental factors that vary over time and location. In particular, the evaluators used in the system should be consistent in the criteria used in evaluating the teacher's performance.

Impact: The system should avoid any serious negative impacts, direct or indirect, on students, teachers, parents, or the community, and if possible have a positive impact on teaching and learning. Formative evaluations of teaching are intended to *improve* the quality of teaching, and therefore, its effectiveness, while summative evaluations are primarily intended to *evaluate* the quality of teaching and have less of a direct focus on improvement. However, the summative evaluation system should certainly avoid having any negative effect on the overall quality of teaching.

In loosely coupled systems—characterized by weak agreement on goals, decentralized control of curriculum and teaching practices, less agreement on the standards of good teaching, and weak norms around improvement—summative teacher evaluation may be more formal and standardized than formative evaluations. If good and effective definitions of teaching are not tightly connected (and agreed to), there can be a great deal of space between messages sent by the formative and summative systems, with the two systems sending conflicting messages (especially if different people with different views of "good teaching" are sending the messages).

Teacher evaluation in loosely coupled systems will need to coordinate carefully between formative and summative systems with particular attention to the fairness, reliability/consistency, and impact of their teacher evaluation systems. In contrast, teacher evaluation in countries that are more tightly coupled—that have stronger agreement on schooling goals, centralized curriculum and teaching standards, more agreement on good teaching, and strong local norms around improvement—will need somewhat less attention on fairness, reliability/consistency, and impact because other aspects of those countries' schooling and professional learning contexts are already coordinated.

2.7 Conclusions

Schools play central roles in their communities and must be understood in their social contexts. In every society, the goals of schooling inform the organization of school systems, their curricula and norms, their conceptions of good teaching, and their approaches to the evaluation of teaching.

We have drawn on Weick's (1976) distinction between tightly coupled and loosely coupled systems as a framework for describing some of the ways in which educational systems function within their societies' social–historical contexts. Following Weick's distinction, we note that country contexts are shaped by social–historical forces and the coupling of systems within those contexts. Approaches to evaluating and improving teaching can therefore be aligned on a continuum, with tightly coupled systems at one end and loosely coupled systems at the other.

In tightly coupled systems, schooling is organized within a centralized framework. The goals, curricula, textbooks, and many teaching practices are centrally delineated, and therefore consistent across the system. Tightly coupled systems also tend to enjoy strong agreement on what counts as good teaching and to share norms about how to achieve it. Schools and the teachers within schools tend to have high levels of collective responsibility for implementing the agreed-upon teaching practices and outcomes.

In contrast, loosely coupled systems tend to be more decentralized, with many more-or-less autonomous units nested within the educational system, each having substantial control over its goals and over what is taught and how it is taught. The curricula, textbooks, and teaching practices tend to vary across schools, and across classes within schools. As a result, in loosely coupled systems there is relatively little consensus about what counts as "good teaching" and, at best, weak norms about how to achieve it.

Paradoxically perhaps, the shared goals of schooling and shared norms for teaching that characterize tightly coupled educational systems allow for a relatively decentralized system for the evaluation and improvement of teaching. With general agreement on the goals and norms of teaching, both formative and summative evaluations of teaching can rely on the shared goals and norms. In loosely coupled systems, it is more difficult to implement evaluation systems based on shared norms and shared conceptions of "good teaching," because there is, at best, limited agreement on what counts as "good teaching." As a result, loosely coupled systems tend to rely on summative evaluation systems based on outcome measures like student test scores that are used to implement accountability criteria for teachers.

2.7.1 Good Teaching and Effective Teaching

An important distinction can be drawn between good practice and effective practice in a profession like teaching. Teaching practice is effective if it achieves its goal or goals, and it is not effective if it does not achieve its goal. Good teaching can be defined at a more general level, in terms of teaching practices that are known to be generally effective (based on research and experience). Good teaching is expected to be effective in most cases, but not in all cases.

The *standards of practice* for teaching depend on the knowledge base for the profession and specify how to deal with the tasks and situations that arise in practice. The standards of practice define the goals and scope of practice and include research-based guidance on determining the nature of the task at hand, general approaches to achieving the desired outcomes, and the specific practices that have been shown to be effective in achieving these outcomes. A particular teaching performance can be judged in terms of whether the performance is consistent with the standards of practice.

Standards of practice for teaching can provide a working definition of "good teaching" and can serve as the basis for evaluations of the quality of teaching, but only if there is general agreement on the *standards of practice*. For the link between "good teaching" and "effective teaching" to be generally accepted, it is necessary to have agreement on the goals of schooling and on effective ways to achieve those goals.

Tightly coupled educational systems tend to have high agreement on what constitutes good teaching, and therefore, on standards of practice for teaching. As a result, samples of teaching can be evaluated in terms of this shared definition of good teaching practice, as specified by the *standards of practice*, with confidence that good teaching thus defined will also be effective in most cases. Loosely coupled systems tend to have much less agreement about what constitutes good teaching, and as a result, confidence in the link between any particular set of teaching practices and student outcomes will be weaker. As a result, in loosely coupled systems, judgments about the quality of teaching based on standards of practice may need to be supplemented by direct evidence of student outcomes in order to be convincing to administrators and the public.

2.7.2 Formative and Summative Evaluations of Teaching

The distinction between formative and summative evaluations of teaching is one of purpose. *Formative evaluations of teaching* provide feedback to teachers that is designed to support improvement while *summative evaluations of teaching* evaluate the teacher's performance in order to make decisions about the teacher. Both kinds of evaluations are designed to improve the overall effectiveness of teaching, but they involve very different theories of action. Formative evaluations of teaching are designed to improve the effectiveness of current teachers. Summative evaluations are designed to improve the overall effectiveness of the teaching staff by rewarding the best teachers and encouraging less-effective teachers to improve their performance. In extreme cases, a summative evaluation of a teacher's performance may result in an administrative decision to remove the teacher from the classroom.

Formative evaluations of teaching are necessarily local. They are generally based on direct observations of teaching and provide feedback intended to help the teacher improve their teaching. Decisions about how the teacher can improve their performance are usually made jointly by the teacher and the evaluator.

Summative evaluations of teaching are likely to be more centralized and to focus on decisions about the teacher and apply to the teacher's general level of performance. They may be based on direct observations of teaching performances, but can also be based on less direct indicators of performance such as VAMs. They are designed to help administrators make decisions about the teacher, and decisions based on the evaluation results are generally more formal and bureaucratic than formative evaluations. Summative evaluation systems that rely on outcome measures (e.g., VAMs) are referred to as "accountability systems."

For tightly coupled systems with high levels of central coordination and collective responsibility for teaching processes and outcomes, formative evaluation and summative evaluation may be integrated in a common system. In loosely coupled educational systems, with little central coordination and a high level of individual responsibility for teaching practice and accountability, the distinction between formative and summative evaluation is likely to be explicitly defined, and the two kinds of evaluation are likely to be implemented separately.

In tightly coupled educational systems, with strong norms for good teaching, formative evaluation tends to be the main concern in teacher evaluation systems, and summative evaluations may play a minor role. In loosely coupled systems, summative evaluations generally play a larger role and may be designed to maintain some degree of control over parts of the system (e.g., teachers, schools).

2.7.3 The Theory of Use for Evaluations of Teaching

A *theory of use* for a teacher evaluation system specifies how results will be used. The main goal for teacher evaluation systems, formative and summative, is to improve the quality of teaching, and thereby, student outcomes, but the two kinds of evaluations do this in different ways. Formative evaluation operates principally through a feedback to the teacher and guidance mechanism, while summative evaluation emphasizes an overall evaluation of the teacher's performance mechanism. The theory of use specifies how the formative or summative evaluation achieves its goal through its main mechanism.

The theory of use for formative evaluations is fairly straightforward. An experienced teaching professional observes the teacher's performance on various criteria associated with specific standards of practice. The results are then shared with the teacher, and the strengths and weaknesses of the performance are reviewed. This process can be assumed to lead to a better understanding on the part of the teacher of their teaching practice, particularly in terms of aspects that could be improved. Given this insight, the teacher and evaluator could then plan some actions that would lead to improved teaching performance.

The theory of use for a summative evaluation system is less straightforward. In principle, teachers could use the results of summative evaluations in the same way that they use the results of formative evaluations, but in practice, the nature of summative results may make using them for improvement challenging. Generally, the results of the summative evaluation are used to make decisions about the teacher's role in the system. Teachers who are identified as performing very well may be given additional responsibility or a promotion. Teachers who are identified as performing badly may be sanctioned or fired. Theories of action for summative evaluations of teaching focus on increasing the impact of the best teachers and limiting the impact of the worst teachers, but they tend to have very limited impact on the performance of most teachers.

The coordination between formative and summative evaluations can be a problem in loosely coupled systems, in which there is relatively little agreement on the definition of good teaching. The two systems can send different messages (especially if there are different people sending the messages and/or are differentially trusted by the teacher), thus causing confusion about goals, norms, and practices within schools.

2.7.4 Final Remarks

Because of their use in supporting high-stakes decisions, summative evaluation systems need to be more formal and systematic than formative systems. The summative evaluation system needs to be transparent, to be fair, and to be perceived to be fair by stakeholders.

In a tightly coupled system, with strong consensus on the nature of good teaching, direct observations of teacher performance can provide a solid basis for both formative and summative evaluation. Teacher evaluations in more tightly coupled, centralized systems have more agreement on the nature of good teaching, and stronger collective and professional standards of practice for teaching. As a result, a well-designed formative evaluation involving observations of teaching and other kinds of inputs (e.g., parent and student questionnaires) might serve both formative and summative purposes fairly well in most cases, with the more formal accountability mechanisms reserved for teachers who are not meeting minimal standards of effectiveness.

Loosely coupled systems have less agreement on the standards of practice for good teaching, and therefore, summative teacher evaluations will tend to be more formal and standardized, with accountability playing a major role in these evaluations. The processes and criteria being used in the one-on-one evaluations that are central to formative uses of these observations tend to be unique from case to case, and therefore, their implications for summative evaluations are not easily communicated to third parties, especially skeptical third parties. So, the outcomes of formative evaluations of teaching may be considered questionable to political leaders and the public in loosely coupled systems.

This limitation can be ameliorated by documenting the processes and outcomes of the formative interactions in some detail. In addition, more standardized observations of performance, with trained, outside evaluators using standardized observation protocols may also improve trust in formative evaluations. Derived measures of student learning might be useful as checks on the formative system because they can be standardized across schools. They are useful in identifying schools that need additional help in promoting good teaching. This kind of formative evaluation system, with audits, would be highly transparent, and assuming that the auditors were well trained, it could promote the effectiveness of the formative evaluation system.

The main purpose of summative evaluations of teaching is to satisfy the need to demonstrate publicly that the schools are functioning effectively. A well-documented formative evaluation system can provide assurance that the schools are achieving the goals assigned to them by the community and therefore can also contribute to this public demonstration.

The answer to the question about how formative and summative teacher evaluation should be structured is complex. The country context and its embedded systems will determine how evaluation systems might best be constructed and used for the improvement and monitoring of teaching.

References

Adnot, M., Dee, T., Katz, V., & Wyckoff, J. (2017). Teacher turnover, teacher quality, and student achievement in DCPS. *Educational Evaluation and Policy Analysis, 39*(1), 54–76. https://doi.org/10.3102/0162373716663646

Ball, D. L., & Forzani, F. M. (2011). Building a common core for learning to teach: And connecting professional learning to practice. *American Educator, 35*(2), 17.

Bell, C. A., Castellano, K. E., & Klieme, E.. (2020). Classroom management. *Global teaching insights study policy report* (Chap. 3). Paris: OECD. https://www.oecd-ilibrary.org/education/global-teaching-insights_20d6f36b-en

Bell, C. A., White, R. S., & White, M. E. (2018). *A systems view of California's teacher education pipeline*. Retrieved from Stanford, CA: https://gettingdowntofacts.com/publications/systems-view-californias-teacher-education-pipeline

Black, P., & William, D. (2003). 'In praise of educational research': Formative assessment. *British Educational Research Journal, 29*(5), 623–637. https://doi.org/10.1080/0141192032000133721

Blömeke, S., König, J., & Felbrich, A. (2009). Middle-school education in Germany. In *An international look at educating young adolescents* (pp. 255–286).

Bloom, B. S., Hasting, T., & Madaus, G. (1971). *Handbook on formative and summative evaluation of student learning*. McGraw-Hill.

Boyd, D. J., Grossman, P. L., Lankford, H., Loeb, S., & Wyckoff, J. (2009). Teacher preparation and student achievement. *Educational Evaluation and Policy Analysis, 31*(4), 416–440. https://doi.org/10.3102/0162373709353129

Briggs, D. (2016). Commentary: Can Campbell's law be mitigated? In H. Braun (Ed.), *Challenges to measurement in an era of accountability* (1st ed, pp. 178–190). Routledge. https://doi.org/10.4324/9780203781302

Chazan, D., Herbst, P. G., & Clark, L. M. (2016). Research on the teaching of mathematics: A call to theorize the role of society and schooling in mathematics instruction. In D. H. Gitomer & C. A. Bell (Eds.), *Handbook of research on teaching* (5th ed., pp. 1039–1098). American Educational Research Association.

Cochran-Smith, M., Villegas, A. M., Abrams, L. W., Chávez-Moreno, L. C., Mills, T., & Stern, R. (2016). Research on teacher preparation: Charting the landscape of a sprawling field. In D. H. Gitomer & C. A. Bell (Eds.), *Handbook of research on teaching* (5th ed., pp. 439–548). American Educational Research Association.

Council of Chief State School Officers. (2013, April). Interstate teacher assessment and support consortium. In *TASC model core teaching standards and learning progressions for teachers 1.0: A resource for ongoing teacher development*. Washington, DC: Author.

Crooks, T. J. (1988). The impact of classroom evaluation practices on students. *Review of Educational Research, 58*(4), 438–481. https://doi.org/10.3102/00346543058004438

Danielson, C. (2007). *Enhancing professional practice: A framework*. ASCD.

Darling-Hammond, L., & Hyler, M. H. (2013). The role of performance assessment in developing teaching as a profession. *Rethinking Schools, 27*(4), 10–15. https://rethinkingschools.org/articles/the-role-of-performance-assessment-in-developing-teaching-as-a-profession/

Darling-Hammond, L. (2021). Defining teaching quality around the world. *European Journal of Teacher Education, 44*(3), 295–308. https://doi.org/10.1080/02619768.2021.1919080

Doan, S. & Mihaly, K. (2021) Regression analysis. *Global teaching insights study technical report* (Chap. 23). Paris: OECD. https://www.oecd.org/education/school/GTI-TechReport-Chapter23.pdf

Fenstermacher, G. D., & Richardson, V. (2005). On making determinations of quality in teaching. *Teachers College Record, 107*(1), 186–213. www.tcrecord.org/Content.asp?ContentId=11694

Fernandez, C. (2002). Learning from Japanese approaches to professional development: The case of lesson study. *Journal of Teacher Education, 53*(5), 393–405. https://doi.org/10.1177/002248702237394

Georgia Department of Education. (2018). *Redesigned college and career ready performance index*. https://www.gadoe.org/Curriculum-Instruction-and-Assessment/Accountability/Documents/Resdesigned%20CCRPI%20Support%20Documents/Redesigned%20CCRPI%20Overview%20011918.pdf

Gitomer, D. H., & Zisk, R. C. (2015). Knowing what teachers know. *Review of Research in Education, 39*(1), 1–53. https://doi.org/10.3102/0091732X14557001

Goodwin, A. L., & Low, E. L. (2021). Rethinking conceptualisations of teacher quality in Singapore and Hong Kong: A comparative analysis. *European Journal of Teacher Education, 44*(2), 365–382. https://doi.org/10.1080/02619768.2021.1913117

Governor's Office of Student Achievement. (2017). *Beating the odds overview*. State of Georgia. https://gosa.georgia.gov/sites/gosa.georgia.gov/files/related_files/site_page/Beating%20the%20Odds%20Two-Page%20Summary%2010272017.pdf

Hallinger, P., Heck, R. H., & Murphy, J. (2014). Teacher evaluation and school improvement: An analysis of the evidence. *Educational Assessment, Evaluation and Accountability, 26*(1), 5–28. https://doi.org/10.1007/s11092-013-9179-5

Hansen, M., & Chu, T. (2014). What we think matters most in evaluating teachers may not be so important: Surprising lessons from redesigning an educator evaluation system. Retrieved from Washington, D.C.: https://www.air.org/sites/default/files/downloads/report/What-We-Think-Matters-Redesigning-an-Educator-Evaluation-System-Dec-2014.pdf

Hill, H. C. (2009). Fixing teacher professional development. *Phi Delta Kappan, 90*(7), 470–476. https://doi.org/10.1177/003172170909000705

Individuals With Disabilities Education Act, 20 U.S.C. § 1400 (2004).

James, J., & Wyckoff, J. H. (2020). Teacher evaluation and teacher turnover in equilibrium: Evidence from DC public schools. *AERA Open, 6*(2). https://doi.org/10.1177/2332858420932235

Jensen, B. Downing, P., & Clark, A. (2017). Preparing to lead: Shanghai continuing professional development. Case studies for school leadership development programs in high-performing education systems. National Center on Education and the Economy: Washington, DC. http://ncee.org/wp-content/uploads/2017/09/PreparingtoLeadSingapore092617.pdf

King, J. A., & Alkin, M. C. (2019). The centrality of use: Theories of evaluation use and influence and thoughts on the first 50 years of use research. *American Journal of Evaluation, 40*(3), 431–458. https://doi.org/10.1177/1098214018796328

Labaree, D. F. (1997). Public goods, private goods: The American struggle over educational goals. *American Educational Research Journal, 34*(1), 39–81. https://doi.org/10.3102/00028312034001039

Lampert, M. (2001). *Teaching problems and the problems of teaching*. Yale University Press.

Lortie, D. C. (1975). Schoolteacher: A sociological study. University of Chicago Press.

Martinez, F., Taut, S., & Schaaf, K. (2016). Classroom observation for evaluating and improving teaching: An international perspective. *Studies in Educational Evaluation, 49*, 15–29. https://doi.org/10.1016/j.stueduc.2016.03.002

National Council on Teacher Quality (NCTQ) (n.d.). Teacher prep review. Retrieved on August 1, 2021, from https://www.nctq.org/review

NBPTS. (2012). Early childhood generalist standards (3rd ed.). Retrieved from https://www.nbpts.org/wp-content/uploads/EC-GEN.pdf

OECD. (2020). *Global teaching insights study policy report*. Paris: OECD. https://www.oecd-ilibrary.org/education/global-teaching-insights_20d6f36b-en

Paine, L., Blömeke, S., & Aydarova, O. (2016). Teachers and teaching in the context of globalization. In D. H. Gitomer & C. A. Bell (Eds.), *Handbook of research on teaching* (5th ed., pp. 717–786). American Educational Research Association.

Papay, J. P., Taylor, E. S., Tyler, J. H., & Laski, M. E. (2020). Learning job skills from colleagues at work: Evidence from a field experiment using teacher performance data. *American Economic Journal: Economic Policy, 12*(1), 359–388.

Peterson, P. E., Woessmann, L., Hanushek, E. A., & Lastra-Anadón, C. X. (2011). Globally challenged: Are US students ready to compete? The latest on each state's international standing in math and reading. *PEPG Report* (No. 11-03). Program on Education Policy and Governance, Harvard University.

Purcell-Gates, V., Duke, N., Stouffer, J. (2016). Teaching literacy: Reading. In D. H. Gitomer & C. A. Bell (Eds.), *Handbook of research on teaching* (5th ed., pp.1217–1268). American Educational Research Association.

Rowen, B., & Raudenbush, S.W. (2016). Teacher evaluation in American schools. In D. H. Gitomer & C. A. Bell (Eds.), *Handbook of research on teaching* (5th ed., pp.1159–1216). American Educational Research Association.

Sahlberg, P. (2007). Education policies for raising student learning: The Finnish approach. *Journal of Education Policy, 22*(2), 147–171. https://doi.org/10.1080/02680930601158919

Sass, T. R., Hannaway, J., Xu, Z., Figlio, D. N., & Feng, L. (2012). Value added of teachers in high-poverty schools and lower poverty schools. *Journal of Urban Economics, 72*(2), 104–122. https://doi.org/10.1016/j.jue.2012.04.004

Scriven, M. (1967). The methodology of evaluation. In R. W. Tyler, R. M. Gagne, & M. Scriven (Eds.), *Perspectives of curriculum evaluation* (Vo1. 1, pp. 39–83). Rand McNally.

Singapore Ministry of Education (MOE). (n.d.). Our teachers. Retrieved on July 31, 2021, from https://www.moe.gov.sg/education-in-sg/our-teachers

Tee Ng, P. (2008). The phases and paradoxes of educational quality assurance: The case of the Singapore education system. *Quality Assurance in Education, 16*(2), 112–125. https://doi.org/10.1108/09684880810868402

Weick, K. (1976). Educational organizations as loosely-coupled systems. *Administrative Science Quarterly, 21*(1), 1–19. https://doi.org/10.2307/2391875

Windschitl, M. & Calabrese Barton, A. (2016). Rigor and equity by design: Locating a set of core teaching practices for the science education community. In D. H. Gitomer & C. A. Bell (Eds.), *Handbook of research on teaching* (5th ed., pp.1099–1158). American Educational Research Association.

Chapter 3
Evaluating Teacher Performance and Teaching Effectiveness: Conceptual and Methodological Considerations

María Paz Fernández and José Felipe Martínez

Abstract Educational theory inextricably links teachers to student learning, as the key factor mediating educational policies and student experiences in the classroom, with research consistently showing a relationship between a range of teacher and classroom variables that exert an important influence on student outcomes. This chapter highlights the key conceptual and methodological issues involved in the evaluation of teaching and teachers, with particular focus on the distinction between the concepts of *performance* and *effectiveness*. It considers the implications of assumptions and choices around *why* the evaluation is conducted, *what* is evaluated, and *how* it is evaluated, presenting a range of methods to collect data on performance and effectiveness. Additionally, we analyze issues related to the reliability and validity of resulting inferences about teacher performance or effectiveness and the implications for policy and practice. Finally, the distinctions and commonalities in evaluating performance and effectiveness in practice are exemplified through the presentation of different models of teacher evaluation.

3.1 Introduction

Teaching and learning are the central constructs of the educational process. Educational theory inextricably links teachers to student learning, as the key factor mediating educational policies and student experiences in the classroom. Educational research supports this notion empirically, consistently showing a relationship between a range of teacher and classroom variables that exert an important influence on student outcomes (e.g., Baker et al., 2010; Brophy & Goode, 1986; Darling-Hammond, 2000; Kane et al., 2013; Rivkin et al., 2005; Tucker & Stronge, 2005).

M. P. Fernández (✉) · J. F. Martínez
UCLA, Los Ángeles, CA, USA
e-mail: mpfernandez@g.ucla.edu

J. F. Martínez
e-mail: jfmtz@g.ucla.edu

J. Manzi et al. (eds.), *Teacher Evaluation Around the World*,
Teacher Education, Learning Innovation and Accountability,
https://doi.org/10.1007/978-3-031-13639-9_3

The general assumption that an improvement in teaching will lead to an improvement in learning (Goldhaber & Anthony, 2007; Hallinger et al., 2014; Hattie, 2009) underlies most teacher evaluation systems—which increasingly are been used as a policy mechanism to trigger and guide efforts to improve teacher practices and consequently student outcomes. The earliest teacher evaluation efforts initially centered on accrediting teacher qualifications, with a focus on knowledge, credentials, experience, and personal characteristics (Martínez Rizo, 2015). The underlying assumption (and often the explicit claim) was that recruiting more talented individuals or improving the qualifications of those already in the workforce would lead to better educational outcomes for students (Porter et al., 2001). As a result, most educational systems nowadays require teachers to obtain some kind of formal teaching credential or certification and/or demonstrate basic knowledge of the content they will teach. However, mounting evidence shows that static indicators of teacher *qualifications* or experience do not sufficiently explain the large variations in student achievement observed in many countries across the world (Harris & Sass, 2009). This has led to a more recent policy shift toward assessing in more detail teacher practices, or more generally the work teachers do inside and outside the classroom, and the impact these practices have on students' learning and other outcomes.

Importantly, while teaching practices and student learning outcomes are closely linked conceptually and empirically, they are also clearly distinct constructs. However, these terms are not always defined or used consistently in the literature or in educational policy. Public reports often explicitly or inadvertently conflate concepts like teacher qualifications, teacher practice, instructional quality, educational experiences, or opportunity to learn and further combine them with outcomes like student test scores, learning trajectories, etc. The resulting *constructs* are often vaguely defined and inconsistently used and may not provide a robust foundation for developing assessment instrument procedures and associated improvement processes. A sample of the literature exemplifies this conceptual inconsistency; different systems may equate student test score gains with teacher (or teaching) *impact* (Rothstein, 2016), *success* (Corcoran, 2010), *growth* (Ehlert et al., 2014), *quality* (Sass, 2008) (Hanushek & Rivkin, 2010), *performance* (Guarino et al., 2012), or *effectiveness* (Glazerman et al., 2010). Numerous researchers have warned about the dangers of reifying the empirical link between teaching and learning, arguing that student test scores cannot capture many key aspects of the broader construct of interest (Baker et al., 2010; Darling-Hammond, 2015). Teacher evaluation from this perspective is a complex undertaking that must consider in appropriate context basic qualifications, experience, and knowledge on one hand, but also the contents taught, the interactions with students around this content, and other aspects of the work of teachers beyond the classroom comprised in a rich definition of the construct *teaching* (e.g., non-academic support, administrative duties, relationships with parents, professional development, mediation with administrators, and so forth).

This chapter highlights the key conceptual and methodological issues involved in the evaluation of teaching and teachers, with particular focus on the distinction between the concepts of *performance* (the work of teachers, broadly defined) and *effectiveness* (the impact teachers can have on relevant student outcomes). The content will focus mainly on the experience in the United States because this is the country in which the debate related to these concepts has taken place the most. We consider the implications of assumptions and choices around *why* the evaluation is conducted, *what* is evaluated, and *how* it is evaluated. More specifically, summative or formative uses and purposes for the evaluation; key constructs, frameworks, and standards underlying the evaluation; and technical properties of methods and measures used to evaluate the key constructs. Ultimately, our focus is on the reliability and validity of resulting inferences about teacher performance or effectiveness and implications for policy and practice (Baker et al., 2010; Kane & Staiger, 2012; National Research Council, 2010).

3.1.1 Purposes and Consequences

The first question posed above (*why* to evaluate teachers) distinguishes among two distinct but related and often complementary uses or purposes: One formative aimed at helping teachers improve their practice by providing feedback on their performance. A second, summative purpose evaluates the teacher over some period of teaching with the goal of making decisions about the teacher (e.g., salary, tenure, dismissal, etc.). Notably, while these purposes require different processes and methods, in reality most evaluation systems seek to balance both formative and summative purposes and structures at least in paper (Bell & Kane, this volume; Wise et al., 1985).

The consequences or stakes of teacher evaluation can vary considerably across different systems. On one end, a purely *formative* system may seek to identify areas of teaching strength and weakness and offers teachers appropriate assistance, professional development, and resources to improve their practice. As accountability models imported from the private sector have made its way into education, high-stakes teacher evaluation has adopted practices such as systems of rewards or sanctions tied to improvements in student learning, including performance pay or salary adjustments, or termination of teachers who do not meet a certain criterion (e.g., Goodman & Turner, 2013; Hanushek et al., 1999; Yuan et al., 2013).

3.2 Constructs, Standards, and Frameworks

Educational accountability systems are shifting focus away from models that center on static markers of teacher qualification, to more closely assessing teacher practices or on-the-job performance (Goe, 2007). This focus requires a robust definition and operationalization of the key constructs involved and their components. As noted earlier, everyday usage of concepts like quality, competence, practice, performance, success, or effectiveness often belies their conceptual richness and the distinctions among them. It can incorrectly suggest that they are exchangeable, or at least that they have widely accepted, consistent definitions on the literature. Historically, accountability reform efforts in education focused on a set of student learning constructs as the key outcomes to improve. Standards like the common core (CC) or next generation science standards (NGSS) operationalize these constructs in terms of key contents and ideas that students should learn in each grade and may further specify what learning of varying levels of depth *looks like*. More recently, many teacher evaluation and accountability systems focus on (or additionally comprise) a set of teacher performance constructs, which are in turn operationalized and scaled into teaching standards and frameworks—these may or may not closely align to the learning standards above.

Teaching and teacher performance are complex multidimensional constructs, comprising a variety of types of knowledge, skill, attitudes, and dispositions (Darling-Hammond, 2006; Kennedy, 2008; Muijs, 2006). In a highly influential paper, Shulman (1987) enumerated the different categories of teacher knowledge, including knowledge of content, curriculum, and pedagogy, but also pedagogical content knowledge (PCK), knowledge of learners, knowledge of educational contexts, and knowledge of educational ends, purposes, and values. This multidimensional definition makes clear that teachers are always expected to know not only the content they teach, but also the most appropriate pedagogical practices and their students' needs and context. Similarly, Bransford et al. (2005) outlined three key constructs: knowledge of learners and how they learn; conceptions of curriculum content and goals; and understanding of teaching "in light of the content and learners to be taught" (p. 10). The authors emphasize that teaching, like other professions, has a social calling and a corpus of academic knowledge that has identified "systematic and principled aspects of effective teaching," supported by verifiable evidence, but also aspects related to tradition, precedent, and experience (p. 12). Reynolds (1992) outlined the competencies, understandings, and personality characteristics expected from teachers to complete the tasks of teaching. For example, the teacher should know the individual students' abilities in order to engage them effectively and also have patience with students who have trouble understanding the subject matter.

Analyzing the historical evolution of the assessment of teacher knowledge in the United States, Gitomer and Zisk (2015) identified four models of increasing proximity to teacher practice, each with different underlying premises, and associated approaches to assessment. The American Council on Education's (ACE) development of the National Teacher Examination (NTE) in the 1940s represents the first model: teachers as educated professionals, which posits merely that teachers should possess a minimum level of *intelligence, culture*, and professional preparation. The second (teacher as a content knowledge professional) grew out of concepts in cognitive psychology which "emphasized the importance of domain-specific knowledge in the acquisition of skill" (p. 38) along with concerns about how the US educational system was preparing students to compete in a globalized economy, as captured by the publication of *A Nation at Risk* (National Commission on Excellence in Education, 1983). The third model (teacher as content knowledge for teaching professional) comprises Shulman's original PCK (Shulman, 1987) and its later adaptation into *content knowledge for teaching* (CKT) (Ball et al., 2008). The most recent model conceptualizes teaching as a knowledge-rich professional practice, and teachers as "learning specialists," with primary emphasis on the application of situated knowledge to inform classroom practice (see also Guerriero, 2018).

Importantly, some authors are skeptical that "the knowledge base for teacher education is developed enough to embody in explicit standards for practice" (Stecher & Kirby, 2004, p. 6). Nevertheless, while there are no mandatory teaching standards at the national level in the United States, standards have been developed for use in many different contexts. The National Board for Professional Teaching Standards were originally published in 1989 as a "guiding framework for every teacher's development of their practice" (National Board for Professional Teaching Standards, 2016, p. 42). Teachers who voluntarily choose to become Board-Certified are expected to demonstrate that their practice meets five *core propositions*.[1] Similarly, the Interstate Teacher Assessment and Support Consortium (InTASC) developed a set of ten Core Teaching Standards outlining what "teachers should know and be able to do to ensure every PK-12 student reaches the goal of being ready to enter college or the workforce in today's world" (Council of Chief State School Officers, 2013, p. 3).[2] InTASC standards outline expected performance, essential knowledge, and critical dispositions for teachers, which have been adopted in many US states in a number of contexts, and most recently and prominently the edTPA (Educative Teacher Performance Assessment) used in dozens of states for initial teacher certification (California Commission on Teacher Credentialing, 2009; Sato, 2014).

The Danielson Framework for Teaching (Danielson, 2013) comprises four domains and 22 components of teaching. While not explicitly presented as standards, the framework operationalizes these components of teaching in terms of expected competencies and behaviors along a developmental continuum (from unsatisfactory to distinguished). The FFT has influenced or been adapted into teaching frameworks and standards that are the basis for teacher evaluation in a great number of districts in the United States and in other countries around the world. For example, the FFT is the basis for teacher evaluation systems such as the ones used in Chile (see Sun in this same volume), Peru (see Espinoza & Miranda in this same volume), New York City, and Quebec, Canada (OECD, 2013).[3]

The Australian Professional Standards for Teachers established in 2011 (revised in 2018) define seven standards, grouped into three domains of teaching (professional knowledge, professional practice, and professional engagement), which outline the expected capabilities at four stages of the teaching career (Australian Institute for Teaching & School Leadership, 2018). The Australian teaching standards also outline

[1] Subject knowledge; commitment to student learning; monitoring and managing student learning; reflecting around and learning about their own practice; and membership in learning communities.

[2] Learner development; learning differences; learning environments; content knowledge; application of content; assessment, planning for instruction; instructional strategies; professional learning and ethical practice; and leadership and collaboration.

[3] In 2020, guidelines for remote teaching were issued for the FFT, which focus on components that are thought to be most relevant for online learning and remote instruction (The Danielson Group, 2020).

the expected competencies for each level of the teaching career, associated with the educator's experience and mastery of the profession. Teachers begin as Graduate after they completed their initial training and can then progress to Proficient when they show they have achieved the seven standards. The next two levels (Highly Accomplished and Lead) are experienced teachers who work collaboratively and can be examples for others in the profession (Australian Institute for Teaching & School Leadership, 2018).

In contrast to Australia, the Teachers' Standards in England are not associated to specific stages of the teaching career and apply to almost all educators regardless of their experience. These Standards, which came into effect in 2012, are divided into two parts and outline the behaviors teachers should exhibit. Part 1 refers to teaching, stating that teachers should "act with honesty and integrity; have strong subject knowledge (…); forge positive professional relationships; and work with parents in the best interests of their pupils" (Department for Education, England, 2013, p. 10). Part 2 outlines the behaviors and attitudes related to teacher's personal and professional conduct, expecting them to "demonstrate consistently high standards of personal and professional conduct" (Department for Education, England, 2013, p. 14).

In contrast to the general frameworks and standards presented above, others are subject-specific and meant to be applied to a particular content area. A range of examples exist, including the Ambitious Science Teaching framework (Windschitl et al., 2018) and the mathematical quality of instruction (MQI) framework (Hill, et al., 2008; Hill et al., 2012; for more examples see, e.g., Bell et al., 2020; Connecticut State Department of Education, 2010; Kloser, 2014; Maine Department of Education, 2012; National Council of Teachers in Mathematics, 2000).

While most subject-specific frameworks focus on math or science, some examples may be found in the language disciplines, for example, the PLATO framework (Protocol for Language Arts Teaching Observation) for effective literacy instruction in English (Grossman et al., 2013). Frameworks can also refer to specific age groups and grades, like the Children's Learning Opportunities in Early Childhood and Elementary Classrooms (CLASS) framework (Hamre & Pianta, 2007).[4] Finally, while frameworks created in the United States and western countries typically focus on classroom behaviors and technical aspects of pedagogy, other international frameworks aim more broadly at *teacher* characteristics, competencies, and even professional and personal profiles (see, e.g., the Singapore Teaching Competency Model, which emphasizes teachers' identity as professionals charged with goals like *nurturing the whole child*, *winning hearts and minds,* or *acting in the student's interest*; Martinez et al., 2016a, 2016b).

[4] The area of emotional support encompasses the dimensions of classroom climate, teacher sensitivity, and regard for student perspectives, while classroom organization includes behavior management, productivity, and instructional learning format. Finally, instructional support is operationalized into concept development, quality of feedback, and language modeling.

3.3 Evaluating Teacher Performance and Teaching Effectiveness

Teacher *performance* evaluation or assessment aims to monitor and judge aspects of instruction and broader professional practice deemed essential or important by a system or key stakeholders. The evaluation entails collecting evidence of classroom instructional practices conducive to student learning, and others seen as important for the daily work of teachers (e.g., collaborating with colleagues or school leadership, engaging with parents and the community, etc.; Goe et al., 2008). It seeks to use approaches and methods that reflect the complexity of teaching—and more generally, teacher on-the-job performance. Authentic, contextualized information and evidence contribute to the real and perceived validity of an evaluation system and can help improve adoption and lessen distrust and resistance (Hamilton, 2005). This is also critical in cases where the evaluation is intended to support formative or improvement goals and for helping teacher education programs promote key skills and practices in teacher candidates (Darling-Hammond, 2008).

By contrast, evaluation of teaching *effectiveness* typically shifts the focus from *inputs* (e.g., teacher qualifications) and *processes* (e.g., teaching) to specific *outcomes* (e.g., student learning as captured by their scores on a standardized test (Meyer, 1996). *Effectiveness* is consequently defined as the extent of change or improvement on student learning outcomes that can be attributed to the teacher or "a teacher's ability to produce higher than expected gains in student test scores" (Goe et al., 2008, p. 5). Standardized tests are relatively easier to collect and less expensive to implement than other outcome and process measures (Cohen, 1995), and to proponents they promise consistent and *valid* comparisons across students and teachers (Papay, 2012). Advances in technology and statistics have made it easier to collect, connect, and analyze longitudinal data in new ways, particularly to create classroom- or teacher-level aggregates reflected changes in student achievement. The highest profile example of this type of approach was the Measures of Effective Teaching (MET) study, and systems of teacher evaluation inspired by it (Bill & Melinda Gates Foundation, 2010), which explicitly define *effective* teachers as those whose students exhibit more *growth* in standardized test scores (and less frequently other types of outcomes). The resulting evaluation systems are summative in spirit and rely on incentive theory, assigning monetary or other rewards and penalties for high and low effectiveness teachers, respectively (Cohen, 1995). Notably, successful teaching here is reflective of individual traits and effort, rather than "a set of learned professional competencies acquired over the course of a career" (Elmore, 1996, p. 16); while other approaches and methods are sometimes used to assess teaching (e.g., observation protocols, student surveys), these indicators are considered useful or valid mainly or exclusively insofar as they are predictive of student achievement outcomes or growth (Kane et al., 2013).

3.4 Methods and Instruments to Assess Teaching Performance and Teaching Effectiveness

3.4.1 General Considerations: Validity and Reliability

Early in the twentieth century, researchers had already identified the challenges and issues related to the *scientific* study of teaching, including selecting among many potential teaching-related constructs, occasions or instances of these constructs in classrooms, and approaches or methods to collect evidence of these constructs, each with particular strengths and weaknesses (Muijs, 2006). The key considerations from a measurement perspective are validity (the degree to which the evidence collected supports a particular inference, interpretation, or use, see AERA, APA & NCME, 2014) and reliability (the extent to which an instrument produces consistent measures of a construct across replications of a measurement procedure). The process of validation entails collecting evidence to support a proposed interpretation or use, with different kinds of evidence typically needed to support different interpretations and uses (AERA, APA, & NCME, 2014; Kane, 2006). Similarly, investigation of reliability in the case of measures of instruction ideally entails assessing the extent of measurement error from a variety of sources (e.g., raters or observers, occasions of measurement, tasks, and dimensions). The role of occasion error is particularly prominent in this context, as instruction is expected to fluctuate across contents, units, days, or even parts of lessons or days—both according to plan, and for unexpected reasons.

Validity and reliability requirements are also tightly linked to the consequences of the evaluation. High-stake evaluations (usually associated with summative purposes) may have more stringent methodological requirements, to ensure that the data used to make the decisions is adequately measuring teachers' practices. This generates additional concerns especially in the case of large-scale teacher evaluation systems (with large numbers of teachers), as the demands for methodological rigor need to be weighed against practical constraints (e.g., feasibility and cost).

A variety of methods and instruments have been developed to measure constructs related to teaching performance and effectiveness as evidence for evaluating teachers (Goe et al., 2008). Each method has different characteristics and properties and combinations of strengths and weaknesses in relation to validity, reliability, and feasibility for particular purposes and in a particular context. The following section provides an overview of a cross section of the most widely used methods and sources of evidence used to measure teaching performance and effectiveness.

3.4.2 Measures of Teaching Performance

Supervisor ratings. Information on teacher practice can be collected through ratings from individuals who supervise teachers, which can include school administrator or personnel from local or national educational agencies, researchers, or outside evaluators. These evaluations are the most common component of teacher evaluation systems in the United States, with evidence collected using a variety of specific approaches (e.g., formal or informal observations; interviews; document review), more or less structured and systematic depending on the goals and stakes (Stodolsky, 1990). The stakes can vary widely within and between systems, from formative uses focused on providing information to teachers on how to improve their practice to higher stakes uses that include decisions related to hiring or promotion (Goe et al., 2008). In general, principals are assumed to have enough contextualized knowledge about teaching performance, and studies have shown adequate reliability and positive correlation between principal ratings of teachers and student achievement (see e.g., Harris et al., 2014; Medley & Coker, 1987). At the same time, questions have been raised about subjectivity, leniency (Hamilton, 2005), reliability (Weisberg et al., 2009), and formative value, since supervisors often lack necessary substantive knowledge, particularly in higher grades (Goldstein & Noguera, 2006).

Peer evaluations. Peer ratings are attractive in teacher evaluation, because colleagues have extensive first-hand information of the knowledge and expertise required in classroom instruction and also the challenges and limitations teachers commonly face. Peer evaluation models such as *Peer Assistance and Review* (PAR)[5] rely on experienced coaches distinguished for their excellence in teaching and mentoring to provide full-time support to incoming and struggling veteran teachers. Some studies have shown that school districts that have implemented PAR have had positive results on retaining novice teachers and dismissing underperforming ones (Goldstein & Noguera, 2006; Johnson & Fiarman, 2012).

However, some evidence indicates that the benefits of peer evaluation accrue only when the evaluated and evaluator have "equivalent in assignment, training, experience, perspective and information about the setting for the practice under review" (Peterson, 1995, p. 100), which constrains the range of application and potentially its feasibility. Other potential issues with peer review include resistance to give negative evaluations to peer teachers, especially colleagues in the same school. Furthermore, these evaluations may lack the necessary credibility within teachers if there is no clear evidence of the evaluator's expertise, leading to no changes in teacher practice (Johnson & Fiarman, 2012).

[5] In these models, teachers who have been identified for their excellence in teaching and mentoring are chosen as coaches to provide support to new teachers as well as experienced colleagues who may require help. Coaches are also responsible for the teachers' formal personnel evaluations. Typically, coaches do not work in a single school, but are matched with teachers from different schools according to grade level or subject area.

Classroom observation. Observations are the most commonly used instruments for teacher evaluation and development all over the world (Bell et al., 2019; Gitomer & Zisk, 2015). In the most basic sense, this approach involves the systematic observation of live or pre-recorded lessons, during which a rater uses an observation protocol, rubric, or rating instrument to systematically register and/or assess teacher practice along a certain continuum or set of categories. Observation enjoys high face validity and has historically been seen as the Gold Standard for measuring instruction, providing direct evidence of teaching as it happens in classrooms, which can best help identify areas for improvement and professional development (Pianta & Hamre, 2009).

Classroom observation instruments can be classified as requiring low or high inference or level of subjective judgment from the rater about the teaching practices they are observing. Low inference refers to actions that observers can readily observe and record, reporting their volume or frequency (e.g., number of times students raise their hand or that the teacher asks question to all students). High-inference measures, on the other hand, require observers to assess instructional practice in terms of various *qualities* or dimensions related to specific constructs (e.g., the teacher asks high-order questions to students) (Wragg, 1999). Most widely known and used observation instruments are high-inference measures (e.g., CLASS, TALIS Video, FFT), each of which defines and operationalizes a set of distinct but related constructs of classroom practice and a continuum of quality to assess them (Martinez & Fernandez, 2019). Subject general observation instruments include the FFT (Danielson, 2013) and CLASS (Pianta et al., 2007), while examples of subject-specific instruments are the ones used in the video study of TALIS in math (OECD, 2020), PLATO in English (Grossman et al., 2013), and RTOP in science (Sawada et al., 2002).

Systematic study of observation measures in the context of teacher evaluation is far from conclusive (Martinez et al., 2016a, 2016b). In general, high-inference observation tends to show lower reliability (Muijs, 2006) or require observers to receive more intensive and expensive training to achieve appropriate levels of reliability (Bill & Melinda Gates). As was mentioned earlier, instruction can vary considerably over time, and therefore, reliability improves when teachers are observed on several occasions (albeit at increased cost). Nevertheless, observation measures have generally lower reliability and precision than traditional self-report and other standardized instruments that do not involve human judgment. Even with rigorous training and certification, high levels of reliability require several raters and occasions (Bill & Melinda Gates). Recent studies highlight these challenges faced in using even the best-known observation rubrics to support inferences and decisions involving individual teachers (Kane et al., 2011). Additional concerns relate to the potential effects the observer may have on the teacher being observed and whether the observed occasion is a representation of the teacher's typical practice or is best conceived as a high watermark (Muijs, 2006).

Teacher surveys and logs. Teacher self-reports of their practice inside or outside the classroom can range from a simple checklist of easily observable behaviors to sets of questions aimed at measuring more qualitative multidimensional constructs. Surveys can be used to study and monitor a wide range of teaching practices at scale and also to assess teachers' dispositions, attitudes, and self-efficacy, in addition to encouraging teachers' self-reflection on their practice (Goe et al., 2008). Surveys comprise a number of items (most of them close-ended) intended to measure one or several aspect or constructs of instruction and teacher practice. An advantage of teacher surveys is their cost-efficiency compared to other instruments (e.g., classroom observation), as they allow to collect data on large numbers of teachers at a relatively low cost and burden to educators. According to Mullens (1995), large-scale surveys are most useful for monitoring four areas of teacher practice: general pedagogy, professional development, instructional materials and technology, and topical coverage within courses. An example of a large-scale survey of teacher practice is the OECD Teaching and Learning International Survey (TALIS), which in 2018 included a sample of 260,000 teachers across 48 countries and economies (OECD, 2019). The Trends in International Mathematics and Science Study (TIMSS), aimed at measuring students' achievement, also collected information on teachers' beliefs and practices in 64 countries in 2019 (Mullis et al., 2020).

Disadvantages of surveys include memory error and social desirability bias, whereby teachers' responses do not reflect real practices or beliefs if they believe these would make them appear in a negative light. These disadvantages can be especially problematic when surveys are used in high-stakes teacher evaluation, such as the self-evaluation Chilean case (see Sun in this same volume). There is also evidence that teacher responses in questionnaires may not match well their instructional practice as recorded by more *authentic* measures based on classroom observations (Muijs, 2006). There are also concerns that teachers may interpret the concepts and aspects of practice in the survey different than researchers and from each other (Ball & Rowan, 2004; Mullens, 1995). For example, survey answers from two teachers may indicate they "always emphasize higher-order skills" (a 5 in a 5-point scale); but these responses may mask substantial differences across teachers, which may over- or underreport the actual frequencies (intentionally or by mistake), or have different interpretations of what is meant by *always*, *emphasize,* or *higher order*.

Teacher logs are brief surveys that are administered frequently in some cycle or period to keep a frequent and detailed record of a small number of typically narrower aspects of practice (Rowan & Correnti, 2009). Because teachers report on their practices frequently, logs reduce problems with memory and recall error prevalent present in end of year and other surveys that cover longer spans of time, resulting in better reliability and generalizability—compared to classroom observation logs which typically comprise much broader samples of occasions and offer better coverage and representation of actual practice (Ball & Rowan, 2004; Rowan & Correnti, 2009). Daily reporting of practice also tends to lessen concerns about interpretation, aggregation, and social desirability in teacher reports. Nevertheless, the advantages of specificity and frequency come at the cost of more nuanced representation of interactions between teachers and students; some researchers argue logs are only suited to studying the *amount* of content taught as opposed to *how* content was taught (Matsumura et al., 2008).

Artifacts and portfolios. Artifacts and portfolios have been used extensively in teacher induction and certification (Martinez et al., 2012). Teachers compile and typically annotate or contextualize a collection of materials and artifacts meant to illustrate their work inside the classroom. Examples of classroom artifacts include lesson plans, assignments, samples of student work, readings, and quizzes among many others. While these instruments traditionally relied on physical materials and paper copies, the incorporation of technology into the data collection process in recent years has enabled electronic portfolios that can comprise images, audio, and videos and can be managed through mobile devices (Kloser et al., 2021).

Portfolios are commonly assumed to represent the teacher's exemplary work and not necessarily their everyday instruction (Goe et al., 2008), but can be structured for daily or routine collection and monitoring typical practice and trajectories of instruction (Martinez et al., 2012). Advantages of artifacts and portfolios lie on their coverage (compared to teacher surveys) and cost (typically lower than observations), as well as strong face validity among teachers and educators, who believe these are an authentic reflection of key aspects of instruction grounded on tangible materials, and present an adequate picture of their instructional practice (Goe et al., 2008). Portfolios can be used to assess important aspects of teaching practice with reliability comparable to observations and other measures that involve human judgment (Stecher et al., 2005). Moreover, portfolio collection requires a strong cognitive commitment from teachers, which makes them valuable learning tools that encourage reflection on instructional practice (Shulman, 1998). Several large-scale teacher evaluation systems over the world are making use of portfolios as part of the instruments to gather information on teacher practices. Prominent examples of portfolios in the United States include the NBPTS certification of excellence, which requires teachers to present a comprehensive structured collection of classroom artifacts and reflections covering lessons and units across a span of months of instruction. At the other end of the teaching career path, the edTPA portfolio (Pecheone et al., 2013) is used in dozens of US states for initial teacher certification. Portfolios are also the basis for the National Teacher Evaluation System in Chile (Taut & Sun, 2014).

Disadvantages of portfolios include, on one hand, their inherent limitation in directly reflecting interactive and verbal classroom instruction and, on the other, the very substantial resources they require to develop, administer, collect, and review. Additionally, portfolios can present a considerable burden on teachers who are responsible for gathering the data over time—although when the process is framed within a professional development cycle, this *burden* is instead seen as the bulk of the work conducive to cognitive growth and learning. Finally, recent critiques of the edTPA call into question whether the psychometric properties of portfolio ratings sufficiently reflect the extent of error and thus uncertainty involved in inferences about individual teachers (Gitomer et al., 2019).

Student questionnaires. Students can be seen as one of the main sources of information on what happens inside the classroom, as they are the ones who spend more time in contact with teachers and their instruction throughout their schooling experience. Student surveys can be used to provide feedback to teachers about how their students perceive their practice, to inform school administrators and communities about average teacher practices in the school, to evaluate individual teachers, and to guide professional development (Bill & Melinda Gates Foundation, 2012a; Kuhfeld, 2017). They are increasingly used as a source of information of and for teacher practice and present important advantages over other instruments. Students' scores report individual student experiences more accurately, but aggregated at the classroom level can offer reliable composites of teacher practices that are more strongly related to student achievement than composites obtained from teacher surveys (Ferguson, 2012). Studies have shown that student surveys can be as reliable as teacher surveys (Martinez, 2012) and classroom observations (Bill & Melinda Gates Foundation, 2012a, 2012b).

Nonetheless, student surveys face significant measurement questions related to error in interpretation (especially with younger students), within- and between-level invariance, and treatment of consensus in student reports. These complexities can lead to misleading or unwarranted inferences and limit the value of the information for informing teacher learning (Schweig, 2016). The issue of within-classroom consistency deserves especial attention and is straightforward to illustrate: consider two classrooms with a mean report of 3 on a 5-point scale reflecting the challenge of assessments and quizzes. One classroom could include two groups of students with radically different perceptions: half the students not challenged (1) and the other half rather overwhelmed (5). In the second classroom, there is perfect consensus and all students reported moderate challenge (3). While both teachers receive a report that shows the same average score, this hides different patterns in responses that show students' experiences with their instruction are qualitatively very different. Appropriately reflecting within-classroom variation can thus be crucial for appropriately interpreting student survey data, and noticing differentiated or individualized instruction, or different student experiences or perspectives within the classroom.

Additional concerns focus on young children's ability to report accurately and biases (negative and positive) or inattention with older students. More broadly, students are able to report on their experiences, but are not technically qualified to assess teachers on specific areas of teaching such as curriculum and content knowledge (Goe et al., 2008). Finally, the exact wording of items can affect student responses, as items with different references can have different psychometric properties (e.g., an item worded as "my teacher asks me to read out loud" may not necessarily be interpreted in the same way as "our teacher asks us to read out loud" (Cole et al., 2011).

3.4.3 Measures of Teaching Effectiveness

Student Growth Models with Test Scores. Student achievement measures are commonly used in the United States to assess schools and teachers' effects on students' learning—they have been a staple of school reform in recent decades.

In contrast to performance measures, evaluating teachers based on student achievement places the emphasis on instructional ends (student learning), rather than means (Popham, 1971), so these models seek to determine the growth in students' achievement and attribute this to the school or the teacher. Models based on student's achievement growth effectively assume, first, that student achievement is a more direct indicator of learning than measures of teacher practice and, second, that achievement measures can accurately and validly predict success in higher education, future earnings, and aggregate economic outcomes (Hanushek & Rivkin, 2010). They thus posit that the ultimate evidence of effectiveness lies on the teacher's ability to have an effect on student learning. Indeed, early proponents argued that there was no clear evidence that teacher behavior was a good predictor of student learning, thus calling into question whether performance measures were ever appropriate to assess student learning (Millman, 1981).

Teacher evaluation based on student's achievement scores in standardized tests has been heavily criticized by experts and the broad educational community. It is argued that privileging summative over formative goals teacher evaluation approaches based on student test scores fail to offer detailed evidence necessary to guide teacher reflection and learning, which is ostensibly a fundamental necessary condition for a system that seeks instructional improvement (Amrein-Beardsley, 2008). Test-based accountability more broadly reduces the idea of "good teaching" to improvement on test scores, effectively assuming that all relevant teaching and learning information can be collected through a standardized test (Apple, 2007).

A practical concern with these instruments is their limited reach. Most standardized tests in use today measure content related to mathematics, reading, or, to a lesser extent, science. The focus on mathematics and reading (English) in the United States can be attributed to requirements from NCLB and ESSA that mandated states to test students in these subjects annually in grades 3 through 8 and then once in high school (Every Student Succeeds Act, 2015; U.S. Department of Education, 2001). Estimates suggest that the longitudinal test scores needed to produce student growth measure estimates are simply not available for as many as 50 to 60 percent of teachers across the US—a sobering reminder of the feasibility of this type of approach, even in the USA, the country that relies most extensively on standardized tests across levels of the educational system. Finally, given the high-stakes nature of these tests and the potential consequences for teachers and schools, the incentive is to reduce the hours spent on teaching subjects that will not be assessed through these tests. Evidence shows that this shift is even more pronounced in school districts serving mostly low-income and minority students, which are more at risk of sanctions for their low scores (Baker et al., 2010). Additional issues are associated with validity (e.g., whether the test measures traits that can be influenced by instruction, if the instrument is used for its intended purpose, among others) and instructional sensitivity of the tests themselves (e.g., the instrument's ability to distinguish between strong and weak instruction; Popham, 2007).

Value-Added Models (VAM). The most prominent effort to advance evaluation of teaching effectiveness in the last two decades has been the advent of so-called *Value-Added* models, which rely on students' scores in standardized tests to estimate the individual effects of a teacher on student learning growth by residualizing average students' test score gains, allowing for more precise indicators of effectiveness (Glazerman et al., 2011). The trend of using standardized tests for school assessment increased in the 1980s, with a surge in test scores used for accountability purposes toward the early 1990s (Linn, 2000). This was heightened with the passing of the No Child Left Behind Act in 2001, that stressed accountability and improvement by making schools prove their effectiveness through Adequate Yearly Progress (AYP) reports[6] (U.S. Department of Education, 2001). Since VAM are longitudinal, they can measure students' progress over time while controlling for "all of the factors that contribute to growth in student achievement, including student family, and neighborhood characteristics," isolating the effect of teachers and schools (Meyer, 1996, p. 200; Goe et al., 2008).

Along with the surge of these methods for teacher evaluation has come strong criticism from educational experts, warning both about psychometric limitations, and broader consequences of strong reliance on test scores. In the specific case of VAM, their face validity is questioned, as teachers do not understand the complex underlying statistics and cannot derive useful information for reflecting on their practice (Grossman et al., 2013). The strongest assumptions behind these models are that students' test scores are a product of their teachers' practices (i.e., a causal relationship between instruction and achievement) (Baker, et al., 2010) and that the aggregates computed can in fact reflect a causal effect. In fact, because students are not randomly assigned to teachers, the presence of bias from unmeasured variables affecting the estimates is always a strong possibility (Rubin et al., 2004). Very few studies have been able to conduct random assignment of students to teacher to establish causality with inconclusive results (e.g., Kane & Staiger, 2008; Kane et al., 2013). Baker et al. (2010) raise further concerns about the inadequacy of statistical controls to account for the student's context and the imprecision and instability of the estimates over time, class, and models (see also Darling-Hammond et al., 2012). Estimates are also inconsistent across achievement measures (Lockwood et al., 2007) which would suggest that effectiveness differs for different skills, in which case estimates should be broken down by subscore.

Student Growth Percentiles (SGP). Other teacher evaluation systems employ student growth percentiles to determine teachers' effectiveness, providing a context for a student's current achievement by locating their most recent score in a conditional distribution that depends on their prior achievement scores. In order to use this information for teacher evaluation, the students' percentiles are aggregated, and the teacher's effectiveness is determined against a defined quantity of adequate student growth whose adequacy can be determined through probabilistic (a fixed growth percentile threshold) or growth-to standard methods (the growth percentile necessary

[6] AYPs were defined as a specific amount of yearly progress in standardized test scores a school, district, or state was expected to make in a year.

to reach the desired performance level threshold Betebenner, 2009, 2011; Walsh & Isenberg, 2013).

An important feature of SGPs is that they are based on a normative conceptualization of student growth, in which the student's learning is measured in comparison with their peers, as opposed to the absolute criterion employed in VAM where the amount of growth is represented by a change in scale score points. Therefore, an advantage of SGPs over VAM is that they tend to be more accessible to teachers and school administrators and can be more easily interpreted. Although both growth models rely on complex estimations to determine the student's actual growth, SGPs provide a percentile rank that has intuitive meaning for the public (e.g., an SGP of 78 means that the student demonstrated more growth than 78% of their peers). However, this normative criterion can also be considered a limitation of SGPs, as these measures by themselves do not provide information on whether the student's relative ranking and their growth are determined to be adequate in their particular educational context (Doss, 2019).

Another perceived strength of SGPs is that they "sidestep many of the thorny questions of causal attribution", focusing on descriptions of student growth that can inform discussions about educational quality (Betebenner, 2009, p. 43). Contrary to VAM, SGPs do not require a vertical scale for the pre- and post-tests (both tests do not have to be on the same scale), so the basic requirement is that they measure the same construct. This is believed to be a more realistic constraint, as a vertical scale is a requisite to estimate VAM estimates (Betebenner, 2011).

However, SGPs present other important limitations. When compared with VAM, SGPs are more sensitive to classroom composition, as they typically do not adjust for student characteristics other than prior achievement (e.g., income, special education status, gender, etc.). This explains in part why SGPs do not perform as well as VAM when students are not randomly assigned to teachers, an assumption that tends to hold in real-life educational situations, implying that teachers who have more disadvantaged students in their class will obtain lower SGP scores than other educators (Doss, 2019; Guarino et al., 2015). Furthermore, research has shown that VAM and SPG models provide dissimilar estimates of student growth and, consequently, of teacher effectiveness, since the estimation methods are different (Goldhaber et al., 2014; Kurtz, 2018).

Student Learning Objectives (SLOs). These measures of student growth are defined as a set of goals that measure teachers' progress in achieving a certain student growth target. They differ from other measures of student growth in that they do not rely on students' scores on standardized tests, but are based on learning targets defined by teachers or educator teams. The development of SLOs follows several steps, where the teacher or education team review of standards identifies core concepts and student needs, sets goals for students, monitors student progress, and finally examines outcome data to determine next steps. Teachers are required to collect baseline and trend data from students in order to determine if they are meeting the goals set for the class. Teachers then must gather baseline and follow-up data, which can come from district assessments, student work sample, and units tests, among other sources (Lachlan-Haché et al., 2012a, 2012b).

SLOs are believed to have several advantages over other types of teacher effectiveness assessments. On one hand, they promote reflections around student results and progress, reinforcing good teaching practices, recognizing teachers' expertise, and empowering teachers as participants in their own evaluation process. On another, SLOs can be adapted to different educational contexts, allowing teacher evaluation to adjust to changes in curriculum or assessments. SLOs can also cover any subject and are not bound by the availability of standardized test scores, which tend to be limited to a few areas of knowledge (reading, mathematics, and science; (Lachlan-Haché et al., 2012a, 2012b).

However, SLOs also present several downsides. Although many states require SLOs to be "rigorous and comparable," providing guidance on acceptable measures to evaluate whether the objectives were reached, meeting the requirements of high-quality assessments and comparability across classrooms, schools, and districts, has proven challenging. Additionally, SLOs should ensure that the growth targets are ambitious while remaining attainable, avoiding the pitfall of setting goals that may be too easy to attain and that may not improve students' learning (Lachlan-Haché et al., 2012a, 2012b).

Other student outcomes. Student achievement is not the only outcome used to assess teaching effectiveness, as there is a growing consensus on the importance of non-cognitive measures that capture the range of effects of schools and teachers on students (Goe et al., 2008; Jackson, 2016; Schweig et al., 2018; West, 2016). Non-cognitive outcomes include higher-order skills like social-emotional learning, student communication and collaboration competencies, critical thinking, creativity, interpersonal competencies, and self-management, among a range of others. Recent research suggests that teachers can have a significant impact on on-time grade progression, absences, suspensions, and other proxies for non-cognitive skills (Jackson, 2016). This study also found that teachers whose practice contributes to the improvement of students' behavior are also able to improve longer-run outcomes like SAT-taking or future GPA scores.

Research is also showing teacher effects on *social-emotional learning (SEL) outcomes,* related to "knowledge, skills, and attitudes to develop healthy identities, manage emotions and achieve personal and collective goals, feel and show empathy for others, establish and maintain supportive relationships, and make responsible and caring decisions" (CASEL, 2020, p. 1). The CASEL framework encompasses five areas of SEL competence: self-awareness, self-management, social awareness, relationship skills, and responsible decision-making. To enhance students' social and emotional skills and attitudes, teachers can employ different practices in a developmentally, contextually, and culturally responsive ways, such as cooperative and project-based learning (CASEL, 2020). An example of the use of SEL as a measure of teaching effectiveness is found in Meyer et al. (2019), who use VAM to estimate the magnitude of classroom-level impacts on students' growth in SEL. The study looks at the effects of the four different constructs measured in the CORE Districts (growth mindset, self-efficacy, self-management, and social awareness), assessing the correlation between the SEL measure and achievement scores. The findings indicate that teachers who improve students' academic test performance may not be the same teachers who promote students' SEL, as there is a low correlation between classroom-level growth in SEL and classroom-level growth in ELA or math, although the growth mindset construct showed a moderately strong relationship.

Even though experts agree that non-cognitive outcomes are relevant and can legitimately be used to assess teachers, what we know about them "is extremely limited because the research has not yielded any truly informative information about how we can achieve any outcomes that we want students to learn in school other than achievement" (Good, 2014, p. 31). Good also points to the lack of consensus around the most relevant non-cognitive outcome and the cost and burden of collecting these alternative outcomes.

Other teacher measures. A range of indicators can be used to capture other relevant behaviors, dispositions, and practices of teaching more broadly defined. Examples of teacher behaviors may include simple markers like attendance, recordkeeping, participation in professional development, ethical behavior, professional interactions with the school community, and collaboration with colleagues, among others. Many systems historically relied on these types of indicators as the primary mechanism for assessing teachers, and these original evaluation systems are still in wide operation around the world as the basic infrastructure of teacher evaluation. An example of this is the teacher evaluation system currently used in the Los Angeles Unified School District in the United States, by incorporating *additional professional responsibilities* as one of the standards in their teaching framework. Within this framework, teachers are expected to maintain accurate records (e.g., track students' progress toward identified learning outcomes, manage non-instructional records, submit the records on time); communicate with families (e.g., inform about the instructional program and the student); and demonstrate professionalism (e.g., show ethical conduct, advocate for students; LAUSD, 2021a, 2021b).

3.5 Designs and Systems

In 2019, twenty-two states in the United States required teachers to be evaluated annually, a decrease from 27 states that evaluated teachers annually in 2015 (NCTQ, 2019). Classroom observations are the most common teacher evaluation measure, currently mandated in 36 states (e.g., Florida, Massachusetts, and New Mexico) and optional in another five (e.g., Arizona, Illinois, and Texas). The most widely used teacher observation protocols are the Danielson Framework for Teaching (Danielson, 2013) used in 18 states and the Marzano Causal Teacher Evaluation Model used in 11 states (Marzano & Toth, 2013). Six other states use rubrics developed either locally or externally in alignment to state standards (Close et al., 2020). Similarly, 31 states currently use student surveys for teacher evaluation, but only seven require these measures (e.g., Hawaii, Iowa, and Mississippi). Student surveys are not used for teacher evaluation in twenty states, and only New York currently prohibits their use (NCTQ, 2019). Finally, 34 states require indicators of learning growth based on student standardized test scores as part of their teacher evaluation system, up from only 15 in 2009, but down from the peak of 43 states in 2015. Of the states that require learning growth data, eight allow using other measures such as district assessments, student portfolios, or student learning objectives, instead of the state's standardized test. When it comes to the particular choice of growth model, 15 states use Value-Added models for summative evaluation, while three more report using these types of VAM scores only for formative purposes—e.g., North Carolina discontinued use of VAM scores for high-stakes personnel decisions and instead uses them to drive teacher professional development (Close et al., 2020). Finally, ten states leave the decision to use VAM scores to local education authorities—for example, in Maine, each school district can measure student growth using one of the two models: VAM or SLO indicators. In Texas, districts can select among VAM, SLOs, portfolios, or other measures to assess student growth (Close et al., 2020).

Table 3.1 presents a cross section of notable US and international teacher evaluation systems and summarizes some of their key characteristics. While not representative in any statistical or qualitative sense, the table reflects the great diversity of systems in terms of purposes, contexts, and technical characteristics and their similarities and differences—for more details about each system, refer to the links in the table.

Some systems focus mainly or exclusively on teacher performance, while deemphasizing or excluding effectiveness, either by design or in practice. For example, the Los Angeles Unified School District (the second largest in the United States) developed a Teaching and Learning Framework based on Danielson's (Danielson, 2013) and aligned to the California Standards for the Teaching Profession (LAUSD, 2021a, 2021b). Performance is assessed through classroom observations, teaching artifacts, student surveys, measures like attendance, and participation in professional development, while student test scores are used only as a benchmark for teachers to establish their own performance objectives. The Chilean Teacher Evaluation System is also based on the Danielson Framework (*Marco para la Buena Enseñanza* (MBE); Ministry of Education, Chile, 2008), but organizes evidence of performance in a portfolio comprising classroom artifacts and scores in an observation rubric (from a videotaped lesson), along with peer evaluation, supervisor ratings, and a self-evaluation rubric.

Conversely, in some systems, effectiveness is the central construct of teacher evaluation. For example, in the IMPACT system implemented at the District of Columbia Public Schools, Value-Added scores (IVA) make up 35% of a teacher's overall evaluation, while an additional 15% is assigned to a student growth measure based on SLOs. Similarly, the state of Florida classifies teachers in four levels of performance, but assigns at least 50% of the weight to VAM indicators of teacher effectiveness (S.B. 736, Student Success Act, 2010). Importantly, because student scores are only available for teachers in certain grades and subjects, schools in DC and Florida must rely on alternative assessments for large proportions of teachers—a

Table 3.1 Comparisons across teacher evaluation models

Education system	Measures	Emphasis	Framework	Combination	Purpose	Additional information
Chile	– Classroom observation – Teaching artifacts – Peer assessment – Supervisor ratings	Performance	MBE (Based on FFT)	Weighted (theoretical/policy)	Formative and summative	https://www.docentemas.cl/ (in Spanish)
District of Columbia	– Classroom observation – Student surveys – Supervisor ratings – Growth models (SLOs and/or VAM)	Effectiveness	DCPS essential practices	Weighted (theoretical/policy)	Formative and summative	https://dcps.dc.gov/page/impact-dcps-evaluation-and-feedback-system-school-based-personnel
Florida	– VAM – Classroom observation	Effectiveness	FFT, Marzano, other approved by state authority	Weighted (theoretical/policy)	Formative and summative	https://www.fldoe.org/teaching/performance-evaluation/
Los Angeles	– Classroom observation – Student survey – Student test scores – Teaching artifacts – Other measures	Performance	LAUSD Standards for Teaching (FFT + CA Standards)	Weighted (Locally determined)	Formative and summative	https://achieve.lausd.net/cms/lib08/CA01000043/Centricity/Domain/433/TLF%20Booklet.pdf

(continued)

Table 3.1 (continued)

Education system	Measures	Emphasis	Framework	Combination	Purpose	Additional information
Met Project	– Classroom observation – Growth models (VAM) – Student surveys – Teacher surveys	Hybrid	FFT, CLASS, PLATO, MQI, and UTOP	Weighted (empirical)	Formative (research-oriented)	https://files.eric.ed.gov/fulltext/ED540960.pdf
New York City	– Classroom observation – Growth models (SLOs and/or VAM)	Hybrid	FFT	Conjunctive	Formative and Summative	https://www.uft.org/sites/default/files/attachments/2020-21_Advance_FAQs_FINAL_051721.pdf

reminder of a fundamental data challenge facing systems that center on effectiveness and student test scores (Baker et al., 2010).[7] After lawsuits challenged this practice, the Florida courts explicitly determined that districts can use school aggregates to evaluate individual teachers (Paige, 2020). Both the DC and Florida systems assign the remaining 50% of the weight in the evaluation using observation measures and other indicators of performance, which individual districts are able to select from approved lists.

The systems in Florida and New York City Schools (the largest district in the United States) exemplify the common hybridization or conflation of the two central concepts underlying this chapter, performance and effectiveness. In New York, eight indicators derived from classroom observations are used for summative *performance* assessment, while the remaining fourteen are used exclusively for non-evaluative feedback. Interestingly, the number of observations each year is determined by the teacher's previous ratings—fewer observations for highly *effective* teachers and more for *ineffective* teachers (New York City Department of Education, 2019). While New York also evaluates teachers using measures of student learning, the model de-emphasizes individual accountability based on effectiveness. A committee with administrators and union members identifies measures, target populations (e.g., different subgroups of students at the classroom, grade, or school level), and even the model (e.g., VAM or goal setting around SLOs).

As for the approach for combining measures, a common hybrid approach combines weighting and conjunctive/disjunctive decision rules or tables. For example, in NY, a teacher rated *ineffective* in the performance measure, and *highly effective* in the measure of student learning is overall classified as *developing*. States like Colorado, Louisiana, or Pennsylvania have implemented similar decision tables. Among systems that use compensatory models, theoretical or policy weights are the commonly used (e.g., DCPS, Florida, Chile) but a variety of other approaches exist. A prominent example is the LAUSD system which frees school sites to determine how to combine information across measures (Los Angeles Unified School District, 2019). The Measures of Effective Teaching study, while not an operating system per se, deserves special mention here as the largest ever to measure teacher performance and effectiveness in thousands of classrooms using multiple observation protocols including FFT (Danielson, 2013), CLASS (Hamre & Pianta, 2007), PLATO (Grossman et al., 2013), MQI (Hill et al., 2008), and UTeach (UTOP; Walkington & Marder, 2018), teacher and student surveys (Ferguson, 2012), supervisor ratings, and even a test of pedagogical content knowledge. Researchers assessed how predictive each measure was of teacher value-added estimates based on standardized test scores and tests of higher-order conceptual understanding (Bill & Melinda Gates Foundation, 2010). The study was very influential in the US during the 2000s and 2010s among other things because it is one of few to randomly assign students to teachers to yield clearer causal effects. However, the various measures were found to

[7] Schools can adopt commercially available tests or develop their own, provided these are "rigorous, aligned to content standards, and appropriate for the teacher's classes and students" (District of Columbia Public Schools, 2011, p. 2; Gitomer & Joyce, 2015).

correlate only weakly and inconsistently to VAM scores, and the authors ultimately emphasized the importance of balancing the weights assigned to performance and effectiveness indicators for high-stakes teacher evaluation—effectively signing away the explicit emphasis on empirical weights that was originally at the core of the study design.

Notably, the results of teacher evaluation conducted over the last few years under this great variety of designs and systems are converging in classifying a great majority of teachers in the highest levels of performance. For example, in Florida, 98% of teachers statewide are rated highly effective or effective, with only 0.6% classified as developing and 0.1% unsatisfactory (Florida Department of Education, 2018), and similar proportions are commonly observed elsewhere (see, e.g., Anderson, 2013; Dynarski, 2016; NCTQ, 2017).

Finally, it is important to note that, as is commonplace across the US and internationally, all the systems in the table claim both summative and formative goals and uses of the measures collected. In Chile, for example, teachers classified as basic or unsatisfactory must complete professional development courses and engage in self-reflection and collaborative peer work to address weaknesses identified in the evaluation, but can eventually face dismissal if they continue to underperform (Taut & Sun, 2014). The DC IMPACT system similarly combines summative consequences for teachers (incentives and potentially dismissals) with individual formative feedback on four areas: instructional practice, student achievement, instructional culture, and collaboration.

3.5.1 Combining Measures to Evaluate Teaching Performance and Teaching Effectiveness

The discussion above makes it apparent that multiple instruments and methods are necessary to provide sufficient information to evaluate teacher performance and effectiveness. Indeed, multiple measures provide a more comprehensive image of both performance and effectiveness (Goe & Croft, 2009), as each of the instruments and measures described earlier is well suited to capture some performance or effectiveness constructs (in some context), but limited or ill-suited to capture others. In addition to improved construct coverage, research shows that multiple measures can produce more stable or precise categories to classify teachers (De Pascale, 2012; Steele et al., 2010), limit score inflation (NCTQ, 2015), and reduce incentives for gaming the system (Steele et al., 2010) among others. Perhaps even more importantly, evidence from multiple measures is needed to provide rich, usable feedback to teachers and thus is essential for constructing strong systems of professional development parallel to the evaluation (Baker et al., 2010; Duncan, 2012). This can also help increase of buy-in among stakeholders (Glazerman, et al., 2011) and identify and reduce adverse impact in time (De Corte et al., 2007).

There are three main approaches to combining evidence from different instruments and constructs (Martinez et al., 2016a, 2016b). *Conjunctive* models assess each measure separately and summarize the information using a joint decision rule—e.g., teachers meet the standard if they obtain a rating of *basic* or above in the observation measure and rank in the top 8 deciles in the student survey and student learning outcomes. This reduces false *positives/passes* by requiring adequate performance in each construct or component (e.g., performance and effectiveness). Conversely, *disjunctive* or complementary models require meeting a criteria for only some measures—e.g., score of *basic* or above in at least two of three measures. This reduces false *negatives/fails* and is preferred when some dimensions are more important than others. Finally, *compensatory* models create a single linear composite index synthesizing the information in the measures—this weighted average allowing high performance on one measure to compensate for lower performance on another (Brookhart, 2009). Weights can be set empirically (e.g., factor analysis, regression coefficients) or theoretically (e.g., through stakeholder negotiation).

Each of these models has advantages and drawbacks and can be used to maximize specific properties of the resulting joint inferences (Mihaly et al., 2013). Importantly, they can also lead to different classifications and decisions for individual teachers (Martinez et al., 2016a, 2016b). In this context, Martinez et al., (2016a, 2016b) suggest that balanced theoretical or policy weights have important advantages because they not only offer desirable psychometric properties in terms of composite reliability and consistency over time, but more importantly reflect a broad stakeholder consensus about the importance of different aspects of teaching performance and teacher effectiveness—a potential powerful hortatory instrument for policy adoption and implementation.

3.6 Conclusions and Implications

Educational improvement efforts centered on teacher evaluation are typically conceptualized around two related but distinct targets of assessment: teacher performance or teaching effectiveness. From the discussion presented above, it is apparent that these approaches rely first on a series of assumptions about the nature and components of *teaching*, a very complex multidimensional construct that is often defined inconsistently by educators, researchers, policymakers, and the public. In addition, these efforts and resulting systems involve assumptions and choices around conceptual and methodological aspects involved in assessing this target construct, and also the most impactful policy mechanisms for exerting influence on it, and the people and organizations involved. For example, Kane & Bell (in this same volume) discuss critical points of distinction between teacher evaluation systems conceived primarily for summative or formative goals.

Importantly, many of these considerations go beyond the strictly technical and relate to broader societal and institutional goals and contexts at the national or subnational level—where a broad range of social and political priorities, pressures, and stakeholders typically play a defining role in spearheading, shaping, modifying, and in some cases ending teacher evaluation systems (see Zorrilla & Martinez in this same volume).

In this chapter, we tried to highlight the complexities associated with these assumptions and choices, the subsequent systematic collection of information and evidence to assess what teachers do (teacher performance), and the effect teachers have on specific student outcomes (teaching effectiveness). The former concept relies on models and frameworks that outline the ideal competencies, practices, and attitudes of teachers. The latter focuses on measuring and improving outcomes, and attaching incentives to the evaluation, with the expectation that this will affect instruction. While effectiveness is often linked with summative goals, and performance with formative objectives, the more useful distinction is at the level of individual instruments or measures, which may be more conducive to formative or summative uses. For example, classroom observation protocols tend to be used in formative teacher evaluation, as they are a source of direct evidence of teaching as it happens in classrooms, which can be used to identify areas of improvement and professional learning for teachers. Conversely, Value-Added Models (or similarly, student growth percentiles) are seen as more summative in nature, as they focus on teachers' ability to improve student outcomes and do not directly offer evidence to guide professional learning or improvement. Importantly, most teacher evaluation systems in operation would reject the summative label; even those with a very strong focus on estimating teacher *effectiveness* typically claim (either explicitly or implicitly) to also have formative value or serve formative goals.

To serve these dual objectives, systems typically rely on the use of multiple measures. While the notion that teacher evaluation requires multiple measures is nearly universal, this idea, like teaching, belies great conceptual and methodological complexity. On one hand, as our chapter outlined, instruments and measures have distinct strengths and weaknesses and may be advantageous for different purposes and in different contexts, inevitably presenting substantive, technical, and practical tradeoffs to developers of teacher evaluation systems. Moreover, different ways of combining information derived from these measures rest on different assumptions and can have direct implications for the inferences made about teaching and teachers. Because no approach to combining measures consistently outperforms the others on strictly technical grounds, systems should thus explore the approach that most closely aligns with their goals and that allows to best illuminate the relevant aspects of the *teaching* construct. Perhaps most importantly, the idea of *combining* the measures into a single final score for each teacher implies a loss of information that in principle would seem counter to the more formative goals these systems typically espouse, as information about specific aspects of the multidimensional construct teaching best illuminated by each instrument is blended into a single ostensibly unidimensional measure (Martinez et al., 2016a, 2016b). Instead, systems should aim to *make combined use* of the information provided by multiple measures, to best utilize the full extent and detail of information provided by each one for formative or summative purposes, or both.

There is mounting evidence, including much reflected in other chapters in this volume, that irrespective of whether performance or effectiveness is the main *narrative* focus, the technical rigor of the instruments is not sufficient to sustain teacher evaluation systems—which additionally require thoughtful implementation, explicit and meaningful focus on improving teacher practice or performance on the ground, and realistic consideration of the institutional, policy, and political context. Without these elements in place, there are no psychometric or statistical techniques, either existing or future, that will enable education systems to sustainably and productively evaluate teachers in very large volumes, on a tremendously multidimensional construct, in complex contexts, and for high-stakes purposes.

References

AERA, APA, NCME. (2014). *Standards for educational and psychological testing.* American Educational Research Association.

Amrein-Beardsley, A. (2008). Methodological concerns about the education value-added assessment system. *Educational Researcher, 37*(2), 65–75.

Anderson, J. (2013, March 30). Curious grade for teachers: Nearly all pass. *New York Times.*

Apple, M. W. (2007). Ideological success, educational failure? On the politics of no child left behind. *Journal of Teacher Education, 58*(2), 108–116.

Australian Institute for Teaching and School Leadership. (2018). *Australian Professional Standards for Teachers.* AITSL.

Baker, E. L., Barton, P. E., Darling-Hammond, L., Haertel, E., Ladd, H. F., Linn, R. L., Shepard, L. A., et al. (2010). Problems with the use of student test scores to evaluate teachers. *EPI Briefing Paper* (278).

Ball, D. L., & Rowan, B. (2004). Introduction: Measuring instruction. *The Elementary School Journal, 5*(1), 3–10.

Ball, D. L., Thames, M. H., & Phelps, G. (2008). Content knowledge for teaching: What makes it special? *Journal of Teacher Education, 59*(5), 389–407.

Bell, C. A., Dobbelaer, M. J., Klette, K., & Visscher, A. (2019). Qualities of classroom observation systems. *School Effectiveness and School Improvement, 30*(1), 3–29.

Bell, C. A., Klieme, E., & Praetorius, A.-K. (2020). Conceptualising teaching quality into six domains for the Study. In OECD, *global teaching insights technical report* (pp. 1–24). OECD Publishing.

Betebenner, D. W. (2009). Norm- and criterion-referenced student growth. *Educational Measurement: Issues and Practice, 28*(4), 42–51.

Betebenner, D. W. (2011). *A technical overview of the student growth percentile methodology: Student growth percentiles and percentile growth projections/trajectories.* The National Center for the Improvement of Educational Assessment.

Bill & Melinda Gates Foundation. (2010). *Learning about teaching: Initial findings from the measures of effective teaching project.* Bill & Melinda Gates Foundation.

Bill & Melinda Gates Foundation. (2012a). *Gathering feedback for teaching. Research Paper.* Bill & Melinda Gates Foundation.

Bill & Melinda Gates Foundation. (2012b). *Asking students about teaching. Policy and practice brief.*

Bransford, J., Darling-Hammond, L., & LePage, P. (2005). Introduction. In L. Darling-Hammond & J. Bransford (Eds.), *Preparing teachers for a changing world: What teachers should learn and be able to do* (pp. 1–39). Jossey-Bass.

Brennan, R. L. (2001). Some problems, pitfalls, and paradoxes in educational measurement. *Educational Measurement: Issues and Practice, 20*(4), 6–18.

Brookhart, S. M. (2009). The many meanings of multiple measures. *Education Leadership, 67*(3), 6–12.

Brophy, J., & Goode, T. L. (1986). Teacher behavior and student achievement. In M. C. Wittrock (Ed.), *Handbook of research on teaching* (3rd ed., pp. 328–375). MacMillan.

California Commission on Teacher Credentialing. (2009). *California standards for the teaching profession (CSTP).*

CASEL. (2020). *CASEL'S SEL framework: What are the core competence areas and where are they promoted?* CASEL.

Close, K., Amrein-Beardsley, A., & Collins, C. (2020). Putting teacher evaluation systems on the map: An overview of state's teacher evaluation systems post–every student succeeds act. *Education Policy Analysis Archives, 28*(58), 1–26.

Cohen, D. K. (1995). Rewarding teachers for student performance. In S. Fuhrman, & J. O'Day (Eds.), *Rewards and reforms: Creating educational incentives that work.* Jossey-Bass.

Cole, M. S., Bedeian, A. G., Hirschfeld, R. R., & Vogel, B. (2011). Dispersion-composition models in multilevel research: A data-analytic framework. *Organizational Research Methods, 14*(4), 718–734.

Connecticut State Department of Education. (2010). *Common core of teaching: Foundational skills.* CSDE.

Corcoran, S. P. (2010). *Can teachers be evaluated by their students' test scores? Should they be? The use of value-added measures of teacher effectiveness in policy and practice.* Annenberg Institute for School Reform.

Council of Chief State School Officers. (2013). *InTASC model core teaching standards and learning progressions for teachers 1.0.* CCSO.

Danielson, C. (2013). *The framework for teaching evaluation instrument, 2013 edition.* Danielson group.

Darling-Hammond, L. (2000). Teacher quality and student achievement: A review of state policy evidence. *Education Policy Analysis Archives, 8*(1), 1–44.

Darling-Hammond, L. (2006). Constructing 21st-century teacher education. *Journal of Teacher Education, 57*(3), 300–314.

Darling-Hammond, L. (2008). Reshaping teaching policy, preparation, and practice: Influences of the national board for professional teaching standards. In R. Stake, S. Kushner, L. Ingvarson, & J. Hattie (Eds.), *Assessing teachers for professional certification: The first decade of the national board for professional teaching standards (Advances in Program Evaluation)* (Vol. 11, pp. 25–53). Emerald Group Publishing Limited.

Darling-Hammond, L. (2015). Can value added add value to teacher evaluation? *Educational Researcher, 44*(2), 132–137.

Darling-Hammond, L., Amrein-Beardsley, A., Haertel, E., & Rothstein, J. (2012). Evaluating teacher evaluation. *Phi Delta Kappan, 93*(6), 8–15.

De Corte, W., Lievens, F., & Sackett, P. R. (2007). Combining predictors to achieve optimal tradeoffs between selection quality and adverse impact. *Journal of Applied Psychology, 92*, 1380–1393.

De Pascale, C. (2012). Managing multiple measures. *Principal, 91*(5), 6–10.

Department for Education, England. (2013). *Teachers' standards: Guidance for school leaders, school staff and governing bodies.* DFE.

District of Columbia Public Schools. (2011). *Teacher-assessed student achievement data (TAS) guidance.* DCPS.

Doss, C. J. (2019). *Student growth percentiles 101: Using relative ranks in student test scores to help measure teaching effectiveness.* RAND Corporation.

Duncan, A. (2012, agosto 22). *Change is hard*. Retrieved from US Department of Education: https://www.ed.gov/news/speeches/change-hard

Dynarski, M. (2016). *Teacher observations have been a waste of time and money.* Brookings Institution.

Ehlert, M., Koedel, C., Parsons, E., & Podgursky, M. J. (2014). The sensitivity of value-added estimates to specification adjustments: Evidence from school- and teacher-level models in missouri. *Statistics and Public Policy, 1*(1), 19–27.

Elmore, R. F. (1996). Getting to scale with good educational practice. *Harvard Educational Review, 66*(1), 1–26.

Every Student Succeeds Act, Title I Section 1111(2)(B)(III)(vi) (2015).

Ferguson, R. F. (2012). Can student surveys measure teaching quality? *Phi Delta Kappan, 94*(3), 24–28.

Florida Department of Education. (2018). *2017–18 District educator evaluation ratings.* Retrieved from Archived Statewide District Evaluation Results: http://www.fldoe.org/teaching/performance-evaluation/archive.stml

Gitomer, D. H., & Joyce, J. (2015). *A review of the DC IMPACT teacher evaluation system.* National Research Council.

Gitomer, D. H., & Zisk, R. C. (2015). Knowing what teachers know. *Review of Research in Education, 39*, 1–53.

Gitomer, D. H., Martinez, J. F., Battey, D., & Hyland, N. E. (2019). Assessing the assessment: Evidence of reliability and validity in the edTPA. *American Educational Research Journal, 58*(1), 3–31.

Glazerman, S., Goldhaber, D., Loeb, S., Raudenbush, S., Staiger, D. O., & Whitehurst, G. J. (2011). *Passing muster: Evaluating evaluation systems.* Brown Center on Education Policy at Brookings.

Glazerman, S., Loeb, S., Goldhaber, D., Staiger, D., Raudenbush, S., & Whitehurst, G. (2010). *Evaluating teachers: The important role of value-added.* Brookings Institution.

Goe, L. (2007). *The link between teacher quality and student outcomes: A research synthesis.* National Comprehensive Center for Teacher Quality.

Goe, L., & Croft, A. (2009). *Methods of evaluating teacher effectiveness.* National Comprehensive Center for Teacher Quality.

Goe, L., Bell, C., & Little, O. (2008). *Approaches to evaluating teacher effectiveness: A research synthesis.* National Comprehensive Center for Teacher Quality.

Goldhaber, D., & Anthony, E. (2007). Can teacher quality be effectively assessed? National board certification as a signal of effective teaching. *The Review of Economics and Statistics, 89*(1), 134–150.

Goldhaber, D., Walch, J., & Gabele, B. (2014). Does the model matter? Exploring the relationship between different student achievement-based teacher assessments. *Statistics and Public Policy, 1*(1), 28–39.

Goldstein, J., & Noguera, P. A. (2006). A thoughtful approach to teacher evaluation. *Educational Leadership, 63*(6), 31–37.

Good, T. L. (2014). What do we know about how teachers influence student performance on standardized tests: And why do we know so little about other student outcomes? *Teachers College Record, 116*, 1–41.

Goodman, S. F., & Turner, L. J. (2013). The design of teacher incentive pay and educational outcomes: Evidence from the New York City bonus program. *Journal of Labor Economics, 31*(2), 409–420.

Grossman, P., Loeb, S., Cohen, J., & Wyckoff, J. (2013). Measure for measure: The relationship between measures of instructional practice in middle school English language arts and teachers' value-added scores. *American Journal of Education, 119*, 445–470.

Guarino, C. M., Reckase, M. D., & Wooldridge, J. M. (2012). *Can value-added measures of teacher performance be trusted?* Education Policy Center at Michigan State University.

Guarino, C. M., Reckase, M. D., Stacy, B., & Wooldridge, J. M. (2015). A comparison of student growth percentile and value-added models of teacher performance. *Statistics and Public Policy, 2*(1), 1–11.

Guerriero, S. (2018). *Teachers' pedagogical knowledge and the teaching profession: Background report and project objectives.* OECD Publishing.

Hallinger, P., Heck, R. H., & Murphy, J. (2014). Teacher evaluation and school improvement: An analysis of the evidence. *Educational Assessment, Evaluation and Accountability, 26*(1), 5–28.

Hamilton, L. (2005). Lessons from performance measurement in education. In R. Klitgaard & P. C. Light (Eds.), *High-performance government* (pp. 381–405). RAND Corporation.

Hamre, B. K., & Pianta, R. C. (2007). Learning opportunities in preschool and early elementary classrooms. In R. C. Pianta, M. J. Cox, & K. L. Snow (Eds.), *School readiness & the transition to kindergarten in the era of accountability* (pp. 49–84). Paul H. Brookes Publishing Co.

Hanushek, E. A., & Rivkin, S. G. (2010). *Using value-added measures of teacher quality.* CALDER - Urban Institute.

Hanushek, E. A., Kain, J. F., & Rivkin, S. G. (1999). *Do higher salaries buy better teachers?* NBER Working Paper No. 7082.

Harris, D. N., & Sass, T. R. (2009). *What makes for a good teacher and who can tell?* CALDER working paper.

Harris, D. N., Ingle, W. K., & Rutledge, S. A. (2014). How teacher evaluation methods matter for accountability: a comparative analysis of teacher effectiveness ratings by principals and teacher value-added measures. *American Educational Research Journal, 51*(1), 73–112.

Hattie, J. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. Routledge.

Hill, H. C., Blunk, M., Charalambous, C., Lewis, J., Phelps, G. C., Sleep, L., & Ball, D. L. (2008). Mathematical knowledge for teaching and the mathematical quality of instruction: An exploratory study. *Cognition and Instruction, 26*, 430–511.

Hill, H. C., Charalambous, C. Y., & Kraft, M. A. (2012). When rater reliability is not enough: Teacher observation systems and a case for the generalizability study. *Educational Researcher, 41*(2), 56–64.

Jackson, C. K. (2016). *What do test scores miss? The importance of teacher effects on non-test score outcomes.* NBER.

Johnson, S. M., & Fiarman, S. E. (2012). The potential of peer review. *Educational Leadership, 70*(3), 20–25.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). American Council on Education.

Kane, T. J., & Staiger, D. O. (2008). *Estimating teacher impacts on student achievement: An experimental evaluation.* NBER Working Paper 14607.

Kane, T. J., & Staiger, D. O. (2012). *Gathering feedback for teaching.* Bill & Melinda Gates Foundation. Retrieved from http://k12education.gatesfoundation.org/download/?Num=2680&filename=MET_Gathering_Feedback_Research_Paper1.pdf

Kane, T. J., Taylor, E. S., Tyler, J. H., & Wooten, A. L. (2011, Summer). Evaluating teacher effectiveness: Can classroom observations identify practices that raise achievement? *Education Next* (pp. 55–60).

Kane, T., McCaffrey, D., Miller, T., & Staiger, D. (2013). *Have we identified effective teachers? Validating measures of effective teaching using random assignment. Research Paper.* Bill & Melinda Gates Foundation.

Kennedy, M. M. (2008). Sorting out teacher quality. *Phi Delta Kappan, 90*(1), 59–63.

Kloser, M. (2014). Identifying a core set of science teaching practices: A Delphi expert panel approach. *Journal of Research in Science Teaching, 51*(9), 1185–1217.

Kloser, M., Edelman, A., Floyd, C., Martinez, J. F., Stecher, B., Srinivasan, J., & Lavin, E. (2021). Interrogating practice or show and tell? Using a digital portfolio to anchor a professional learning community of science teachers. *Journal of Science Teacher Education, 32*(2), 210–241.

Kuhfeld, M. (2017). When students grade their teachers: A validity analysis of the tripod student survey. *Educational Assessment, 22*(4), 253–274.

Kurtz, M. D. (2018). Value-added and student growth percentile models: What drives differences in estimated classroom effects? *Statistics and Public Policy, 5*(1), 1–8.

Lachlan-Haché, L., Cushing, E., & Bivona, L. (2012a). *Student learning objectives as measures of educator effectiveness: The basics.* American Institutes for Research.

Lachlan-Haché, L., Cushing, E., & Bivona, L. (2012b). *Student learning objectives: Benefits, challenges, and solutions.* American Institutes for Research.

LAUSD. (2021a, April 3). *History of EDST*. Retrieved from https://achieve.lausd.net/Page/11782#spn-content

LAUSD. (2021b). *Teaching and learning framework.* LAUSD.

Linn, R. L. (2000). Assessments and accountability. *Educational Researcher, 29*(2), 4–16.

Lockwood, J. R., McCaffrey, D. F., Hamilton, L. S., Stecher, B., Le, V.-N., & Martinez, J. F. (2007). The sensitivity of value-added teacher effect estimates to different mathematics achievement measures. *Journal of Educational Measurement, 44*(1), 47–67.

Los Angeles Unified School District. (2019). *2018–2019 EDS final evaluation report for teachers and non-classroom teachers: Administrator handbook.* LAUSD.

Maine Department of Education. (2012). *Common core teaching standards.* MDE.

Martínez Rizo, F. (2015). La evaluación del desempeño docente. Una propuesta para la educación básica en México. In G. Guevara Niebla, M. T. Melendez Irigoyen, F. E. Ramon Castaño, H. Sanchez Perez, & F. Tirado Segura (Eds.), *La evaluación docente en México* (pp. 64–95). INEE-Fondo de Cultura Económica.

Martinez, J. F. (2012). Consequences of omitting the classroom in multilevel models of schooling: An illustration using opportunity to learn and reading achievement. *School Effectiveness and School Improvement, 23*(3), 305–326.

Martinez, J. F., & Fernandez, M. P. (2019). Evaluación docente con indicadores múltiples: Consideraciones conceptuales y metodológicas en torno a la validez. In J. Manzi, M. R. Garcia, & S. Taut (Eds.), *Validez de Evaluaciones Educacionales en Chile y Latinoamérica* (pp. 531–562). Ediciones UC.

Martinez, J. F., Borko, H., & Stecher, B. (2012). Measuring instructional practices in middle school science using classroom artifacts. *Journal for Research in Science Teaching, 49*, 38–67.

Martinez, J. F., Schweig, J., & Goldschmidt, P. (2016a). Approaches for combining multiple measures of teacher performance: Reliability, validity, and implications for evaluation policy. *Educational Evaluation and Policy Analysis, 38*(4), 738–756.

Martinez, J. F., Taut, S., & Schaaf, K. (2016b). Classroom observation for evaluating and improving teaching: An international perspective. *Studies in Educational Evaluation, 49*, 15–29.

Marzano, R. J., & Toth, M. D. (2013). *Teacher evaluation that makes a difference: A new model for teacher growth and student achievement.* ASCD.

Matsumura, L. C., Garnier, H. E., Slater, S. C., & Boston, M. D. (2008). Toward measuring instructional interactions "at-scale." *Educational Assessment, 13*, 267–300.

Medley, D. M., & Coker, H. (1987). The accuracy of principals' judgments of teacher performance. *The Journal of Educational Research, 80*(4), 242–247.

Meyer, R. H. (1996). Value-added indicators of school performance. In E. A. Hanushek & D. W. Jorgenson (Eds.), *Improving America's schools: The role of incentives* (pp. 197–223). The National Academies Press.

Meyer, R., Pier, L., Mader, J., Christian, M., Rice, A., Loeb, S., Hough, H., et al. (2019). *Can we measure classroom supports for social-emotional learning? Applying value-added models to student surveys in the CORE districts.* PACE.

Mihaly, K., McCaffrey, D., Staiger, D., & Lockwood, J. R. (2013). *A composite estimator of effective teaching (MET Project).* The RAND Corporation.

Millman, J. (1981). Student achievement as a measure of teacher competence. In *Handbook of teacher evaluation* (pp. 146–166). Sage.

Ministry of Education, Chile. (2008). *Marco para la Buena Enseñanza.* MINEDUC.

Muijs, D. (2006). Measuring teacher effectiveness: Some methodological reflections. *Educational Research and Evaluation, 12*(1), 53–74.

Mullens, J. E. (1995). *Classroom instructional processes: A review of existing measurement approaches and their applicability for the teacher followup survey.* U.S. Department of Education.

Mullis, I. V., Martin, M. O., Foy, P., Kelly, D. L., & Fishbein, B. (2020). *TIMSS 2019 international results in mathematics and science.* TIMSS & PIRLS International Study Center.

National Board for Professional Teaching Standards. (2016). *What teachers should know and be able to do* (2nd ed.). NBPTS.

National Commission on Excellence in Education. (1983). *A nation at risk: The imperative for educational reform.* U.S. Department of Education.

National Council of Teachers in Mathematics. (2000). *Principles and standards for school mathematics.* NCTM.

National Research Council. (2010). *Preparing teachers: Building evidence for sound policy.* National Academy of Sciences.

NCTQ. (2015). *State teacher policy yearbook: National summary.* National Council on Teacher Quality (NCTQ).

NCTQ. (2017). *Running in place: How New teacher evaluations fail to live up to promises.* NCTQ.

NCTQ. (2019). *State of the states 2019: Teacher & principal evaluation policy.* National Council on Teacher Quality (NCTQ).

New York City Department of Education. (2019). *Advance guide for educators 2019–2020.* NYCDE.

OECD. (2013). *Teachers for the 21st century: Using evaluation to improve teaching.* OECD Publishing.

OECD. (2019). *TALIS 2018 results: Teachers and school leaders as lifelong learners* (Vol. 1). OECD Publishing.

OECD. (2020). *Global teaching insights: A video study of teaching.* OECD Publishing.

Paige, M. (2020). Moving forward while looking back: How can VAM lawsuits guide teacher evaluation policy in the age of ESSA? *Education Policy Analysis Archives, 28*(64), 1–18.

Papay, J. (2012). Refocusing the debate: Assessing the purposes and tools of teacher evaluation. *Harvard Educational Review, 82*(1), 123–141.

Pecheone, R. L., Shear, B., Whittaker, A., & Darling-Hammond, L. (2013). *2013 edTPA field test: Summary report.* SCALE.

Peterson, K. D. (1995). *Teacher evaluation: A comprehensive guide to new directions and practices.* Corwin.

Pianta, R. C., & Hamre, B. K. (2009). Conceptualization, measurement, and improvement of classroom processes: Standardized observation can leverage capacity. *Educational Researcher, 38*(2), 109–119.

Pianta, R. C., La Paro, K. M., & Hamre, B. K. (2007). *Classroom assessment scoring system.* Paul H. Brookes.

Popham, W. J. (1971). Performance tests of teaching proficiency: Rationale, development, and validation. *American Educational Research Journal, 8*(1), 105–117.

Popham, W. J. (2007). Instructional insensitivity of tests: Accountability's dire drawback. *Phi Delta Kappan*, 146–155.

Porter, A., Youngs, P., & Odden, A. (2001). Advances in teacher assessments and their uses. In V. Richardson (Ed.), *Handbook of research on teaching* (4th ed., pp. 259–297). AERA.

Reynolds, A. (1992). Getting to the core of the apple: A theoretical view of the knowledge base of teaching. *Journal of Personnel Evaluation in Education, 6*, 41–55.

Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, schools, and academic achievement. *Econometrica, 73*(2), 417–458.

Rothstein, J. (2016). *Can value-added models identify teachers' impacts?* IRLE—UC Berkeley.

Rowan, B., & Correnti, R. (2009). Studying reading instruction with teacher logs: Lessons from the study of instructional improvement. *Educational Researcher, 38*(2), 120–131.

Rubin, D. B., Stuart, E. A., & Zanutto, E. L. (2004). A potential outcomes view of value-added assessment in education. *Journal of Educational and Behavioral Statistics, 29*(1), 103–116.

S.B. 736, Student Success Act. (2010). *St.* FL.

S.B. 736, Student Success Act, Section 1012.343(3)(a)1 (2010).

Sass, T. R. (2008). *The stability of value-added measures of teacher quality and implications for teacher compensation policy.* CALDER—Urban Institute.

Sato, M. (2014). What is the underlying conception of teaching of the edTPA? *Journal of Teacher Education, 65*(5), 421–434.

Sawada, D., Piburn, M. D., Judson, E., Turley, J., Falconer, K., Benford, R., & Bloom, I. (2002). Measuring reform practices in science and mathematics classrooms: The reformed teaching observation protocol. *School Science and Mathematics, 102*(6), 245–253.

Schweig, J. D. (2016). Moving beyond means: Revealing features of the learning environment by investigating the agreement of student ratings. *Learning Environments Research, 19*(3), 441–462.

Schweig, J., Baker, G., Hamilton, L. S., & Stecher, B. M. (2018). *Building a repository of assessments of interpersonal, intrapersonal, and higher-order cognitive competencies.* RAND Corporation.

Shulman, L. (1998). Teacher portfolios: A theoretical activity. In N. Lyons (Ed.), *With portfolio in hand* (pp. 23–37). Teachers College Press.

Shulman, L. S. (1987). Knowledge and teaching: Foundations of the new reform. *Harvard Educational Review, 57*(1), 1–22.

Stecher, B. M., Wood, A. C., Gilbert, M., Borko, H., Kuffner, K. L., Arnold, S. C., & Dorman, E. H. (2005). *Using classroom artifacts to measure instructional practices in middle school mathematics: A two-state field test (CSE Report 662).* CRESST.

Stecher, B., & Kirby, S. N. (2004). *Organizational improvement and accountability: Lessons for education from other sectors.* RAND Corporation.

Steele, J., Hamilton, L. S., & Stecher, B. M. (2010). *Incorporating student performance measures into teacher evaluation systems.* The RAND Corporation.

Stodolsky, S. S. (1990). Classroom observation. In J. Millman & L. Darling-Hammond (Eds.), *The new handbook of teacher evaluation: Assessing elementary and secondary school teachers* (pp. 175–190). Corwin Press.

Taut, S., & Sun, Y. (2014). The development and implementation of a national, standards-based, multi-method teacher performance assessment system in Chile. *Education Policy Analysis Archives, 22*(71).

The Danielson Group. (2020). *The framework for remote teaching.* The Danielson Group.

Tucker, P. D., & Stronge, J. H. (2005). *Linking teacher evaluation and student learning.* Association for Supervision and Curriculum Development.

U.S. Department of Education. (2001). *No child left behind act (Executive Summary).* U.S. Department of Education.

Walkington, C., & Marder, M. (2018). Using the UTeach observation protocol (UTOP) to understand the quality of mathematics instruction. *ZDM Mathematics Education, 50*, 507–519.

Walsh, E., & Isenberg, E. (2013). *How does a value-added model compare to the colorado growth model?* Mathematica Policy Research.

Weisberg, D., Sexton, S., Mulhern, J., & Keeling, D. (2009). *The widget effect: Our national failure to acknowledge and act on differences in teacher effectiveness.* The New Teacher Project.

West, M. R. (2016). Should non-cognitive skills be included in school accountability systems? Preliminary evidence from California's CORE districts. *Evidence Speaks Reports, 1*(13), 1–7.

Windschitl, M., Thompson, J., & Braaten, M. (2018). *Ambitious science teaching.* Harvard Education Press.

Wise, A. E., Darling-Hammond, L., McLaughlin, M. W., & Bernstein, H. T. (1985). Teacher evaluation: A study of effective practices. *The Elementary School Journal, 86*(1), 60–121.

Wragg, E. C. (1999). *An introduction to classroom observation.* Routledge.

Yuan, K., Le, V., McCaffrey, D. F., Marsh, J. A., Hamilton, L. S., Stecher, B. M., & Springer, M. G. (2013). Incentive pay programs do not affect teacher motivation or reported practices: Results from three randomized studies. *Educational Evaluation and Policy Analysis, 35*(1), 3–22.

Chapter 4
Political Tensions Around Teacher Evaluation

Margarita Zorrilla and Arcelia Martínez

Abstract In recent decades, different countries in Latin America and other regions around the world have established ambitious public policies in educational evaluation to transform educational systems and improve their quality and equity in an effort to ultimately improve student learning. Although in this context the evaluation of teachers has been presented as one of the most important links to transform educational systems, developing this type of evaluation has been marked by various areas of tension, which makes it clear that political factors play a leading role in this activity. In this chapter, different areas of tension that arise in the teacher evaluation process are analyzed to reveal more about the "black box" of a public policy that involves a varied set of actors with different positions on the aims and outcomes of the evaluation. It is argued that these tensions also derive from aspects which encompass the various purposes, types, and consequences of the evaluation, the times during which each actor expects the changes to occur, the lack of governance involved in the evaluation processes, and the ineffective communication about the benefits of teacher evaluation.

During the 2013–2018 period, Margarita Zorrilla served as advisor to the Governing Board of the defunct National Institute for the Evaluation of Education (INEE), where she had an important role and responsibilities in the design and implementation of the teacher evaluation policy in Mexico and other components of the educational reform of those years.

Between 2014 and 2017, Arcelia Martínez worked as a civil servant at the National Institute for Educational Evaluation, where she was in charge of the General Directorate of Guidelines for the Improvement of Education, an area responsible for evaluating policies—including initial training and development of teaching professionals—and the issuance of educational policy recommendations.

M. Zorrilla (✉) · A. Martínez
Universidad Iberoamericana, Ciudad de México, México

A. Martínez
e-mail: arcelia.martinez@ibero.mx

J. Manzi et al. (eds.), *Teacher Evaluation Around the World*,
Teacher Education, Learning Innovation and Accountability,
https://doi.org/10.1007/978-3-031-13639-9_4

4.1 Introduction

> The quality of an educational system cannot exceed the quality of its teachers.
>
> Barber and Mourshed (2007)

The evaluation of teachers has been presented as one of the most important links in transforming the teaching profession, which is why, for at least three decades, various processes and evaluation devices have been implemented to attract the best teachers to the profession and to strengthen their performance and development during their years in service. Nevertheless, teacher evaluation policies, as they are directed at one of the most and best organized unions, can face resistance, especially if these policies threaten the rights and privileges won by the union. The foregoing, coupled with obstacles and tensions of various kinds—technical, financial, administrative, and political, among others—compromises and puts in check the most rigorous evaluation designs, which demands that the actors responsible for such designs and their implementation must dialogue with and arrive at plausible agreements that make it viable.

The effectiveness of teacher evaluation policies requires, therefore, looking at much more than their technical dimension since, as Corrales points out in his analysis (1999), "the success" of teacher evaluation is defined and determined in political terms. It is necessary that the evaluation be, in the first instance, the product of an explicit agreement between the most significant political forces of a country and approved by the legislature in the constitutional and legal reforms that are necessary, but also that a broad consensus is reached among other actors in charge of implementing it, in addition to being accepted by its final recipients: the teachers. In this regard, Corrales affirms that reforms are more likely to overcome political obstacles if they are capable of addressing four aspects: the concentration of costs and the dispersion of benefits; the deficiencies in the degree of commitment of the ministries (the offer); the deficiencies in the degree to which societies demand reforms (demand); and the institutional characteristics that increase the power of the groups that can exercise veto power (such as the teachers' unions) (p. 3).

That said, this chapter analyzes different tensions that arise in the process of teacher evaluation to shed some light on what could be called the "black box"[1] of a public policy[2] that involves a varied set of actors who participate in the design and execution of the evaluation, all of whom may have different positions regarding the aims and consequences of the evaluation. Although this book shows experiences from various countries that exemplify the complexity of teacher evaluation policies,

[1] Easton (1957) uses the metaphor of the "black box" to refer to what we do not always observe in the political system, where positions, interests, and agendas of all the members of the system collide, placing demands and providing support so as to be transformed into legislation and policies.

[2] A public policy is, according to Luis Aguilar, a set (sequence, system, cycle, spiral) of intentional and causal actions that are aimed at achieving objectives considered of value to society or solving problems whose solution is considered of interest or public benefit (2010: 29). In this definition, it is also important to understand public policy as a process, "the basis of which is undoubtedly decision-making, but which implies activities that precede and follow government decisions" (p. 31).

this chapter has as its main reference the Mexican experience framed in the 2013 educational reform. In that teacher evaluation process carried out between 2013 and 2018, a large number of actors converged with different interests, motivations, positions, and responsibilities—elements that gave this process a particularly illustrative stamp to analyze those tensions.

This chapter is structured in four parts. In the first part, the justifications for the policies of teacher evaluation are briefly discussed, including the *why* and the *what for*. In the second part, a basic map of the actors is drawn that shows how these actors converge in the teacher evaluation process. This section looks at the roles these actors play and the distinct positions they represent given the policies in question. Part three presents five tensions that surround the policies of teacher evaluation and how these compromise not only its design but also its implementation and as a consequence the promise for the educational system's improvement. A last section presents a set of final considerations by way of conclusions.

In the Mexican case, it is important to point out that the 2013 evaluation was, for the first time, a compulsory evaluation for all teachers of basic and upper secondary education, which had employment and salary implications; in addition, the design, communication, and implementation of this policy had to be carried out in record time—less than six years—taking into consideration that, in that time, it is estimated that around one and a half million teachers had to be evaluated.

In 2013, a constitutional and legal reform was enacted that was aimed at teacher professional development, defining elements of the teaching career from entering the educational service to retirement or separation. This was called the "Professional Teaching Service (SPD)." Based on a new law, The General Law of Professional Teaching Service, four processes were defined for the teaching profession: entry, promotion, recognition, and permanence for which an evaluation system was developed that permitted decisions to be made about each case. Through these definitions, they sought to install the notion of merit to gain access to a teaching position or a position of leadership, to obtain incentives, and to assure job security. With the educational reform of 2019, the SPD was abrogated, and a new teaching statute was established—The General Law for the Career System for Teachers—in which the teaching evaluation was coordinated with its consequences.

In fact, political tensions were present from the beginning of the changes. Among the weakest factors of the reform, we can mention determination of times that prevented an adequate maturation of the decisions; ineffective communication about the evaluation process; lack of dialogue with the teaching profession; and a complex network of actors that made it difficult to build a system of governance conducive to smooth execution of the actions derived from the policy contained in the Professional Teaching Service (SPD).

The analysis of what happened in the Mexican case with respect to the educational reform of 2013, and, specifically, with the repeal of the SPD, it is still in the process of being analyzed in greater depthneeds, from our experience and involvement as officials in the National Institute for the Evaluation of Education (INEE), organism that disappeared in 2019 with the new educational reform, we can affirm that the political factor—in which the purposes, motives, histories, and interests of those who participated are condensed—was the most important and even decisive in judging the events and results of a policy that ended up being unsuccessful, in the eyes of different social, academic, business, and even current public administration actors.

4.2 The *Why* and the *What for* of Teacher Evaluation

Pressure to improve the quality of learning has been increasing considerably with the reforms of the last decade of the twentieth century. Since the year 2000, when the Program for International Student Assessment (PISA) carried out the first international measurement of the learning of competences in three important areas—language and communication, mathematics and science—this pressure has increased.[3] Today, we also know that the ability of countries to compete in the knowledge economy depends on how they face the growing demand for a high level of knowledge in their citizens, which undoubtedly requires a substantive improvement in the quality of student results as well as an equitable distribution of opportunities to learn. In this context and using PISA results, in 2007 McKinsey and Company conducted the first cross-country comparative study to answer the question: what do high performing and rapidly improving education systems have in common?

The McKinsey Report (Barber & Mourshed, 2008) constituted a watershed, concluding that successful educational systems are based on the quality of their teachers: "getting the right people to become teachers, (1) developing them into effective instructors and, (2) ensuring that the system is able to deliver the best possible instruction for every child" (p. 6). As can be seen in this statement, the quality factor per excellence is teaching and, consequently, the educational professionals who practice it. Then, it would seem that it is not possible to expect good learning results and educational quality if there is no guarantee in the quality of the teaching process for which the teachers themselves are responsible.

For this reason, the educational systems of the Latin American and Caribbean region and other countries around the world prioritized a large part of their efforts on improving the quality of education, and the "professional performance of the teacher" was positioned as a highly influential factor to achieve a significant change in educational outcomes (Cuevas & Tiburcio, 2016; Bruns & Luque, 2014; Ravela, 2012; Guzmán, 2005, Schulmeyer, 2002).

Indeed, since the 1990s, several Latin American countries, including Mexico, began to consider different teacher evaluation policies. Although a good part of the teacher evaluation was linked to the granting of salary incentives, simultaneously, the official discourse oriented the concept of evaluation as a triggering mechanism to improve the quality of teaching and the quality of education overall [4] (Rivas et al.,

[3] According to data from 2018, for that year, 37 OECD member countries and 42 more participated in PISA, including countries and economies such as Shanghai, Hong Kong, and Macao. This data speaks of the growing importance of the search to improve the quality of education.

[4] In Mexico, for example, the experience of teacher evaluation, as we now know it, is new in terms of its relationship to entry into the public educational service, promotion to managerial positions,

2020; Cuevas & Tiburcio, 2016; Santibañez et al., 2007). The McKinsey Report, for its part, greatly influenced the insistent look at the quality of teachers, and based on this, two important questions were added to the road map of educational policies: How to improve the quality of teaching? and How to ensure the effectiveness of the teaching profession? The evaluation of teacher performance was consolidated as an instrument to significantly spur improvement in the quality of teachers and that of student learning (Ravela, 2012). Since then, educational quality and teacher evaluation have been seen as an indissoluble binomial, such that the teaching profession became, if it was not already, a profession that had to be promoted, regulated, and of course, evaluated.

4.3 The Set of Actors in the Teacher Evaluation Policy

In the various educational approaches and reforms promoted, especially since the 1990s, it is very important to identify which actors participate and are influential (Corrales, 1999) in order to dialogue and, sometimes, convince them. In the case of teacher evaluation, we find those who are located in the different levels or orders of government (central, state, and municipal or their equivalents), those who have positions of authority or directors, and those positioned in the old-fashioned bureaucracies who have in-depth knowledge of public administration in different sectors, not just education. In the various teacher evaluation processes, those within the teaching profession and their unions and professional organizations, as well as academics, researchers and/or evaluation specialists, businesspeople, civil society organizations, and mothers and fathers, among others, converge.

With so many actors involved, it is very useful to create a map, an exercise which Silva (2017) argues is necessary to reduce the costs of building relationships of political consensus around public problems and solutions. The process of constructing such a map of relevant actors can help us to better understand what each actor has to gain or lose so as to improve communication, exchange of information, and the possibility of achieving consensus.

With this in mind, we sketch a basic map of the set of actors in the development of teacher evaluation policies. This map is not exhaustive and groups together different actors and circumstances. A more rigorous analysis of political feasibility (Majone, 1992) will require a more detailed and disaggregated description by type of actor[5] that locates the different positions, interests, and strengths of each to be able to

the granting of monetary or other incentives, and the meeting of teacher performance standards as a condition for job permanence. The most important antecedent was the Teaching Career Program (1993–2013) which, although its expressed purpose was to improve the quality of teaching and consequently of learning results, its origin was linked to the salary compensation that was urgent in the teaching profession (Santibáñez et al., 2007).

[5] The map would show levels of commitment, participation, resistance, empowerment, and coalitions, i.e., the power, position, and interaction strategies of the actors, to determine their power, organization, interests, and areas of tension (Silva, 2017).

Table 4.1 Map of relevant actors in the teacher evaluation process (in reference to the Mexican case)

Actors and institutions	Stage of the policy cycle in which they usually participate	Position before the teacher evaluation
Governing bodies of the executive power – Ministry of education – Ministry of finance – Local education authorities	Design, implementation, and evaluation	In favor
Legislature – National congress – Local congress – Parliamentary factions	Establishment of agenda and design	Variable, depending on each political party and its factions
Technical bodies of a public and private nature that coordinate and/or participate in the evaluation – Evaluation institutes – Research centers – Higher education institutions	Design, implementation, and evaluation	In favor
Teachers – Unions and educational associations – Individual teachers	Agenda setting, design, implementation, and, in some cases, evaluation	Variable, since they are not homogeneous actors; usually, the unions are against
Academics working in universities and research centers	Establishment of agenda, design, and evaluation	Variable, since they are not homogeneous actors
Civil society organizations	Establishment of agenda, design, and evaluation	Variable, since they are not homogeneous actors
International organizations	Establishment of agenda, design, and evaluation	In favor

Source Prepared by the authors

look at their veto capacity, neutrality, or support for the exercise of the evaluation. Such a map would be a guide through each stage of the public policy cycle, which according to Aguilar (2010) includes the establishment of agenda, the definition of the public problem, the formulation of the policy, the construction of options to solve the public problem, the decision between options, the communication of the policy, the implementation, and the evaluation (Table 4.1).

4.3.1 Governing Bodies of the Executive Power

Regarding the teaching career in general and teacher evaluation in particular, the ministries of education are the government agency responsible for addressing the definition and content of educational policy, as well as establishing the conditions for the governability and governance of the educational system, through the management and organization of the different government areas. The design and implementation of a comprehensive policy on the evaluation of teachers require, therefore, that the legal standing over the organization that must evaluate teachers be precisely established along with the distribution of responsibilities among the central and local authorities.

In addition to thinking about the old dilemma between centralization *versus* decentralization when deciding who should have which authority or another, it will also be important to keep in mind that government bodies are not merely composed of rules, organizations, and decisions, but are made up of people that besides to exercising their corresponding responsibilities and functions, also imprint their particular point of view on policy. As Merino (2010) points out, public policies always suppose an affirmation of values, "it is on that battlefield where the values that have been selected for a given problem and for the design of the proposed solutions are put into play" (p. 47).

Another body of the Executive Power that plays a key role is the Ministry of Finance or its equivalent because of its responsibility to define the amounts and distribution of the government budget for the execution of programs and public interventions. In fact, negotiations between both ministries (i.e., Ministry of Finance and Ministry of Education) are essential to have the necessary financial resources to operate not only the policies for evaluation but also the consequences of the evaluation with their budgetary implications in terms of salaries, incentives, and in-service training programs among others.

4.3.2 Legislative Branch

In addition to the approach made by government agencies, at the central or local level, other institutions and state actors that are involved in the policy-making process, such as the Legislative Power, must be considered. This branch is responsible for the elaboration and approval of the laws that will frame the teacher evaluation policies. In fact, the scope of the evaluation is usually determined, first, in a constitutional text, which must be approved, as in the case of Mexico, by a qualified majority (two-thirds of Congress) and by at least 17 of the 32 local legislatures.

The participation of Congress in teaching policies is, in fact, a daily exercise, as it is in its functions to review, and if necessary to modify and annually approve the budgets that will be allocated for teacher training and evaluation. Thus, for example, in the case of Mexico, the national Congress, through the House of Representatives, is responsible for annually approving the budget project of the federal (central) government destined for educational overall—and therefore for the evaluation process—which has been prepared by the Ministry of Finance, based on the pre-project that in turn is sent by the Ministry of Education. It should also be noted that Congress is not a homogeneous entity and that partisan factions operate within it that have different views on the objectives and scope of the evaluation, as well as other educational policies.[6]

4.3.3 Technical Bodies of a Public and Private Nature

Due to the experience of several countries in Latin America and other latitudes, the technical dimension of the design and implementation of public policies in education and other areas of social life is essential so that actions planned occur and purposes formulated are achieved. The technical part of the evaluation refers to the design and validation of instruments, the application, review and analysis of results, the production of general and personalized reports, and, ultimately, the guidelines—based on the evaluation—for initial teacher training and for in-service training. The technical component of the evaluation is very important so that when analyzing each specific case, it will be necessary to see who performs each part of the process and if each person responsible has the experience and technical skills to carry out the teacher evaluation. The same must be analyzed whether these components rely on institutions, organizations, or people.

Although with different administrative and financial characteristics, the truth is that today there are a variety of institutions in charge of evaluation, ranging from centers within public or private universities to private centers with mixed financing and specialized units within the ministries or institutions created expressly to carry out the evaluation. In this regard, it should be noted that, since the 1990s, several countries in Latin America have created specialized institutions dedicated to educational evaluation. This is the case of the National Institute for Educational Studies and Research (INEP) in Brazil, the National System for Measurement of the Quality of Education (SIMCE) and the Education Quality Agency (ACE) in Chile, the Colombian Institute for the Evaluation of Education (ICFES) in Colombia, the National Institute of Educational Evaluation (Ineval) in Ecuador, the National Institute for the Evaluation of Education (INEE) in Mexico—the institution was cancelled with the 2019 education reform—and the National Institute of Educational Evaluation (INEEd) in Uruguay.

4.3.4 Teachers

A fundamental actor, although with a different degree of participation and collaboration in the evaluation processes, is the teachers' union, which in most cases acts in

[6] For a more specific analysis of the different positions of the parliamentary factions that made up the so-called Pact for Mexico, which gave rise to the educational reform of 2013, and which included the creation of the Professional Teaching Service as a central component, the work of Oscar Daniel Hernández González is recommended reading (2021).

a coordinated manner, as this guarantees the defense of common causes. In fact, the unions' presence and even participation in decisions will be a function of their own political strength, often determined by the number of their members, but, above all, by their ability to negotiate with those from the governments who lead educational reforms and evaluation policies.

In this regard, Corrales (1999) points out that "compared to other actors who must bear the costs, the teachers' unions enjoy comparative political advantages as pressure groups … therefore, if they turn against the reforms, they can seriously undermine the process" (p. 21). The unions, depending on their relationship with the government apparatus and a series of factors also linked to the logistics and communication regarding the evaluation, will declare they are for or against.

At this point, it is also worth looking at individual teachers, who do not always share the views of the union or the teachers' association. Teacher support for the union, the union cause, and/or the teacher evaluation policy will be based on their age and years of service, the educational level in which they work, and the subjects they teach. An account of the different positions about the evaluation of teacher performance framed in the educational reform of 2013 is available in a study of teachers working in basic and upper secondary education coordinated by the INEE (2016). The study shows—contrary to what some detractors of said reform would affirm—that not all teachers were dissatisfied with the evaluation. Criticism of the process had more to do with the forms and the perception of poor communication regarding the evaluation including the times involved, the logic and meaning of the evaluation, and uncertainties about the effective institutional channels to respond to the questions and doubts the evaluation process generated.

Different views were found among the participating teachers about the relevance of the Performance Evaluation to assess their daily work. At one extreme, there were those who believed that, despite the complications of the process, the evaluation stages do recover aspects of their daily activity in the classrooms (although in some stages more than in others), allowing an assessment of their work as teachers and the capacity to identify from this assessment some possibilities to strengthen it.

At the other extreme, there were teachers who believed that the purpose of evaluating their performance as teachers was not fulfilled because the mechanisms put in place and the instruments used at each stage had a series of problems and inconsistencies that prevented an adequate assessment of their daily work in their classrooms and schools (INEE, 2016).

4.3.5 Academics Working in Universities and Research Centers

Another important group is made up of academics and researchers interested in evaluation issues, and, specifically, in teacher evaluation, who seek to influence for or against policies, from more technical or political positions. In this regard, Weiss

(2016) points out that even in the most objective and dispassionate research, there is an inevitable intrusion of values so that the political and philosophical attitude of researchers is influenced by their theoretical vision, initial assumptions, and methodological preferences and by incomplete explanatory models. Thus, for example, some researchers can be described as "symbolic (technical) analysts" and may even hold positions in the public administration of education. These symbolic analysts, according to Braslavsky and Cosse (1996) "would differ from traditional officials by their awareness of the relationship between knowledge and power, and their conviction that they possess relevant knowledge for the effective exercise of power" (p. 2).

An important consideration about the teachers' unions and the "symbolic analysts" is that, like the other actors, they are not homogeneous ideologically and much less so in their political positions. However, it should be noted that unlike teachers' unions that generally tend to close ranks when government decisions affect their interests, analysts, academics, or researchers behave differently by maintaining and expressing their differences of position either for or against the current policies. Their positions vary among them and depend on various factors such as their position vis-à-vis the decisions of the educational authority, or their closeness to the teachers' union and their union organizations, or their understanding of teacher professional development.

4.3.6 Civil Society Organizations

Civil society organizations (CSOs) are another actor that, through different means, have sought to promote and further advance the quality of education. The set of CSOs is also varied in its composition and in the pursuit of objectives: these actors may consist of organizations of mothers and fathers, of education professionals, or businesspeople who express their interest in the professional development of teachers and therefore of teaching. In the case of Mexico, for example, some civil society organizations such as Suma por la Educación and Mexicanos Primero, financed by businesspeople, played an important role in the 2013 education reform by advancing the idea that it was very important to have teacher evaluations that were mandatory for all and had consequences for their promotion and permanence, in order to improve teachers' pedagogical deficiencies and improve student learning (see Nava & Rueda, 2014, who analyze the role of the media and organized civil society in the construction of the public educational agenda).

Among the social actors, there are also, increasingly, some opinion leaders, who, with the emergence and proliferation of social networks, tend to have an influence, sometimes decisive, in the positioning, criticism, observation, and evolution of teacher evaluation policies. In Mexico, for example, different spaces for observation and analysis of educational policy have emerged, such as México Evalúa, which did a specific follow-up of the 2013 educational reform, or Educación Futura, an education and journalism portal that monitors educational events and offers spaces to voice the opinions of different key actors in the system, among whom are teachers and academics from public and private universities.

4.3.7 International Organizations

The pressure to participate in the knowledge economy and that exerted by international organizations such as the Organization for Economic Cooperation and Development (OECD) and the search to improve equality in educational opportunities have transformed the quality of education into a priority for governments from different latitudes in the world. Consequently, educational policies have been designed and implemented to improve the quality of education, including those that seek to renew or transform the teaching profession through important modifications in recruitment, initial training, and ongoing professional development.

Thus, for example, the OECD report "Improving Schools: Strategies for Action in Mexico" (2010) points out, in its preface, that "it is part of the OECD's efforts to support the reform of the OECD member countries and associated countries," as a "result of the agreement established with the Ministry of Public Education (SEP) to improve the quality and equity of the educational system in Mexico (2008–2010)" (p. 3). The OECD, having under its responsibility the coordination of PISA, has played an important role in promoting the public agenda of teacher evaluation, as well as in advising and conducting studies related to the subject, and thus, it lays out a global strategy, with recommendations to improve, among other things, the quality and potential of teachers "through clear national standards, placing greater emphasis on their training, professional development, selection, hiring and evaluation processes" (OECD, 2010, p. 3).

4.3.8 Actors Matter and… a Lot

Given the above, when looking at the feasibility of a teacher evaluation policy, we conclude that mapping the institutions, various groups, and actors involved in this process is an inescapable and entirely necessary task. Identifying the set of key actors that participate in the different stages of the policy cycle, with different positions, can help to explain and anticipate possible conflicts and/or tensions that impact not only the technical design, but also its operation and management, and even and ultimately, the governance of the educational system itself.

Indeed, Corrales (1999) affirms that the approval and implementation of educational reforms, including the evaluation of teachers, continues to be as difficult as it has always been, while "political obstacles continue to paralyze and distort reform initiatives" (p.4). In the view of this author, it is not only imperative to understand these obstacles, but also to carry out a cost and benefit analysis in the implementation of the policy, as long as the success of the policies (i.e., that they achieve their goals) depends on the extent to which the costs and benefits are concentrated or dispersed.

> Specialists argue that when the costs of a particular policy fall directly and intensely on specific interest groups and its benefits are too widely dispersed, the adoption of such a policy is difficult from a political point of view (see Wilson, 1973, in Corrales, 1999, p.4).

While teacher evaluation policies are an example of dispersed benefits and concentrated costs, it is not surprising that cost-bearing groups, such as central and local bureaucracies, and teachers' unions themselves make people feel their discontent or disapproval. In the following section, we analyze some of the main areas of tension around teacher evaluation policies. Anticipating these can help us consider how best to mitigate the always latent conflict and disagreement.

4.4 Five Areas of Tension Surrounding Teacher Evaluation Policies

Not only do different actors converge in teacher evaluation policies, as already mentioned, but different values and visions of what should and can be done are also in dispute. In this section, we point out that this is an arena full of tensions, which derive from the visions, sometimes conflicting, of the actors involved in its design and implementation, but also from the multiplicity of purposes, types, and consequences of the evaluation, the times the different actors imagine the process will take, the lack of governance in the evaluation processes, and the ineffective communication about the benefits of the evaluation. The five areas of tension mentioned in Table 4.2 are not an exhaustive list, but they do constitute a basis for looking at the possibilities of conflict. In addition, and as will be seen, they are interrelated.

Table 4.2 Areas of tension around teacher evaluation policies

1. The broad set of actors who define and reinterpret the public problem to be solved and the possible solutions
2. The different purposes and types of evaluation and their consequences
3. The timing in which the changes are expected to occur
4. The lack of governance in the evaluation processes
5. The ineffective communication regarding the benefits of the evaluation

Source Authors' elaboration

4.4.1 The First Area of Tension: The Broad Set of Actors Who Define and Reinterpret the Public Problem to Be Solved and the Possible Solutions

Perhaps, one of the most important areas of tension concerns with the conflicting visions of the actors involved in the different stages or life cycle of politics. The diversity of actors involved can be seen in the policy cycle model (Aguilar, 2010; Subirats, 2008). The cycle refers to the different stages in which policies occur, ranging from agenda setting and/or the entry of the public problem in the public and government agenda, to the design of the intervention, its implementation, monitoring, and evaluation.

Understanding the future of a policy, its success, difficulties, and even its failure implies identifying its origin to see in what context it arises, which actors participated in the construction of its objectives, who had the most weight in what was suggested and what was ultimately the "winning" definition of the public problem to be solved. The phase of including the problem in the government's agenda is followed by phases of design and implementation, where new assertions of values are established (Merino, 2010) by the set of actors who participate in the process. Generally, it is not the same actors who participate in one phase or another, and, therefore, there will necessarily be different values at stake. For this reason, the agreements reached in the design phase will face new interpretation processes at the time of policy execution.

Although the specialists are in charge of carrying out the task of design, validation of instruments, elaboration of protocols and rules for the application, as well as the review, issuance and delivery of results, it is the senior officials of the Ministry of Education who have under their responsibility the negotiation and construction of agreements with the teachers' unions and with the officials who manage the public finances.[7] The Ministry also has in its hands the authorization of the budgets to carry out the actions of the evaluation policies.

In addition, not only the Ministry and central administrative bureaucracies participate in the implementation, but many other actors at the local level, such as school supervisors, administrators, and teachers are also involved. In fact, one of the main areas in question on the side of those who operate the evaluation is that the technical design or the objective of the evaluation is often in the hands of those who have little to do with the daily reality of the teachers and the problems they face every day in schools.

Both the design and implementation of policies go through different processes of interpretation and redefinition, as each actor reads and translates them based on their agenda and set of values and beliefs. Therefore, an implementation that is presumed will be relatively successful will need to consider several elements focused on honest and open discussions and analysis between those who designed the policy and those who are responsible for implementing it.

[7] In all these groups, there are those individuals in leadership positions who promote and support the actions of public policy and those who sabotage them or express their open opposition.

4.4.2 The Second Area of Tension: The Different Purposes and Types of Evaluation and Their Consequences

Another area of tension, linked to the previous one, derives from the different purposes, types, and consequences of the evaluation. It is not the same to plan a competitive entrance evaluation to the teaching career choosing those who demonstrate better skills based on a teaching model previously established by the country's educational authority, as it is to carry out an evaluation that grants salary incentives and other benefits.

In the case of Mexico, for example, the new teaching statute of 2013 known as the Professional Teaching Service (SPD) had the declared purpose of creating a system for the professionalization of teachers in which the evaluation and training processes went hand in hand, thus coherently contributing to the improvement of educational quality.[8] However, the establishment of evaluation processes for entry into the teaching career had as one of its starting points the inheritance of teaching positions, so, as Sierra (2017) points out the evaluation (that of entering the service) was actually looking "to end uses and customs regarding the allocation of places that had nothing to do with the suitability of teachers, but rather with patronage-based relationships within the National Union of Education Workers (SNTE)" (pp. 8–9).

In addition, although a major purpose associated in macro with the exercise of teacher evaluation is the professional improvement of teachers, from a development perspective, evaluation should be at the service of improvement, thus leaving an unfinished discussion about what should be the "consequences" of the evaluation. Among the various positions is the one that argues that teacher evaluation must have strong consequences to be taken seriously, the same as in the cases of selection of new teachers, or new university students, or the granting of benefits or incentives to teachers, or the like.

On the opposite side are those who point out that the evaluation of teachers only has the function of improvement and learning. It is at this point that we are, in fact, in a confrontation between what is usually called formative evaluation and summative evaluation. While the first is associated with improvement purposes and is usually seen as an evaluation of soft consequences, the second is linked with strong consequences (for a more complete discussion of the distinction between the formative and summative purposes of teacher evaluations, see the chapter in this same book by Bell and Kane).

In the Mexican case, as proposed by the educational reform of 2013 regarding job separation—the General Law of the Professional Teaching Service indicated that after three opportunities, if the evaluated teachers had an insufficient result in their evaluation, they would be removed from teaching in front of group to occupy, instead, an administrative position. Defining the consequences of teacher evaluation is undoubtedly one of the elements with the greatest political implication; hence, its

[8] In the study carried out by OREALC-UNESCO at the request of the now-defunct INEE in 2016, the dispute over the purposes of the evaluation and the actual implementation of the evaluation can be seen (INEE/OREALC/UNESCO, 2017).

management is a delicate matter, at the risk of losing the fragile balances that some of the leaders may have built. In fact, looking at the reality of the tension created between formative evaluation and summative evaluation leads us to think that governments, in the management of educational evaluation systems, need to work toward building a different culture of evaluation, placing it as a tool for support and an opportunity for improvement, and less as a punishment.[9]

In summary, the teacher evaluation policy will face difficulties to arrive smoothly at effective design and implementation partly because most of the time there are no shared visions about the target in terms of its objectives and purposes or with its consequences. This is not only explained by the various actors that intervene in the different stages of politics, but also by the professional and institutional origin of its designers and implementers (Merino, 2010) and very importantly, according to Sabatier (1988), by the set of beliefs of the groups, which, by way of promoting coalitions, obtain a place on the agenda of discussion.

4.4.3 The Third Area of Tension: The Timing in Which Changes Are Expected to Occur

One more area of tension in teacher evaluation policies derives from the intersection of times of different natures in which changes are expected to occur. Braslavsky and Cosse (1996) point out that one of the main difficulties in policy development and implementation can be explained by the existence of four types of times—political, specialist-professional, bureaucratic, and pedagogical—each with different logics.

Political time according to Braslavsky and Cosse (1996) is what is managed by the national government and in particular by the Ministry of Education. It is related to the purposes of the government, the way in which those purposes are managed, the relations among the political forces acting on stage at a given time, as well as the ability to generate agreements and consensus on specific issues, such as teacher evaluation. Political time stems mainly from external demands such as the electoral calendars of national and even local governments, as well as from the times of the political life of the teachers' unions.

The specialist-professional time, on the other hand, refers to the time for the construction of the knowledge required to offer the foundation, the legitimacy, and the consistency necessary for the execution of actions. This is time in the hands of actors with a more technical and/or scientific profile, who normally have to convince political actors of the viability of a given policy within a specific social and political timeframe.

Bureaucratic time, according to Braslavsky and Cosse (1996), is defined as the path required to comply with all the steps defined by the regulations, without running administrative or legal risks. It is not surprising, then, that the bureaucracies installed

[9] On this subject, reviewing Ravela's comments in Manzi et al. (2011, p. 222) is recommended reading.

in the ministries in times prior to the implementation of new policies tend to show their opposition, in principle, for fear of losing power in the management of the educational system.

Finally, there is pedagogical time, which, according to Braslavsky and Cosse (1996), refers to the time required for the learning journey that the actors must traverse to appropriate a new policy, and even to contribute to its improvement in one or more of its dimensions. Something that should be emphasized is that normally the pedagogical times are not on the horizon of those who design the policy and of the teams that work on its implementation, which in addition to being a factor for the failure of the teacher evaluation policy will result in criticism of the evaluation process in general.

Undoubtedly, the analysis of actors and their times does not refer to an abstract concept, but to people, with professional life projects of professional development, with consequences for teaching and learning in a given educational system. The areas of tension are always present and can lead to constant confrontation resulting in the erosion, sometimes prematurely, of a policy that seeks to promote a challenging and demanding transformation.

4.4.4 The Fourth Area of Tension: The Lack of Governance in the Evaluation Processes

One more area of tension is related to the lack of governance in the evaluation processes. In this regard, it should be noted that in recent decades the idea that governments cannot function alone has become very important, both because of the thinning of the state in its various functions as a result of the crisis of the eighties, and, also, because of the increasing complexity of the public problems to be solved.

Aguilar (2006) defines governance as "the way of distinctively naming the new reality that arises and that concerns the direction or governance of society, but which is a different reality from that of governing by the government alone and includes and integrates the interactions of various actors" (p. 110). In this sense, governance, unlike governability, will denote something more than the mere directive action of the government; it is a less vertical form of government response, where people are willing to negotiate and invest time, based on the recognition that it is necessary for various social and political actors to participate in the design and implementation of the public policies. In governance, the citizen is recovered as a "crucial agent of the governmental environment, whose behaviors and demands in social and political life represent 'opportunities' or 'threats/adversities' for the legitimacy, reliability and effectiveness of the government" (Aguilar, 2006, p. 44).

Regarding the latter, Aguilar (2010) argues that in order to be classified as public, a policy requires, among other things, that the opinion, participation, and co-responsibility of the public citizens be incorporated; that it is to say, transparent and accountable to the public for its actions and results; and also, that it looks out for the public interest and benefit. Beyond the formal definition of what a public policy is, what is clear is that in the face of the complexity of social actions that seek to modify realities, different—less hierarchical—forms of governance are required, based on the idea of a governance that organizes, gathers, listens, and holds all the actors involved in said policy accountable.

In the case of policies related to teachers, which are usually complex and conflictive, the lack of explicit involvement at the invitation of the authority and the inability to reach agreements with the various actors involved can block the implementation of the policies, however well-meaning and technically flawless they may be. For this reason, in recent decades, spaces have been opened not only to political actors linked to local governments and trade union organizations, but also to members of civil society, such as family, academia, and some CSOs—which have demanded spaces for participation and placed different discussions on the public agenda—which undoubtedly makes reaching agreements even more complex. The evaluation, therefore, cannot be done without the participation of these actors, while also as a priority including the teachers themselves who, ultimately, will be the allies or major detractors of the policy.

A teacher evaluation policy based on a true governance scheme should include the accountability of the different actors, governmental and social, in the tasks that seek to improve the educational system. However, the outcome of the policy will depend on the agreements and commitments assumed, and on the follow-up that the civil society actors themselves—academia, *think tanks,* and organizations—can give to the policy, in a kind of virtuous cycle that mediates among the demand for accountability, feedback, and informed accompaniment.

4.4.5 The Fifth Area of Tension: Ineffective Communication Regarding the Benefits of the Evaluation

Finally, we want to end this list of areas of tension that is not exhaustive, nor ordered in degree of importance or chronological sequentially, with one more area of tension, which derives from the poor communication regarding policy. As Martinic (2011: 19) has pointed out: “The formulation of educational reform policies and their implementation constitutes a broad social and communicative process... The process of change takes place in a complex system of relationships in which the actors intervene with their own frames of reference from which they think, define their interests and collective strategies of action1. In these interactions, consensuses, dissents and spaces of uncertainty occur”. Thus, we strongly consider that the success of a teacher evaluation policy not only depends on the ability to listen to the many actors involved and interested in it, and summon them to dialogue, but also on the ability to have an effective communication regarding the benefits of the policy with all of them, mainly with the teachers, who will be affected directly.

In this exchange—of listening and communication—the justifying arguments that are used to support the teacher evaluation still play an important role in the elaboration (and success) of the policies, because, as Majone (1995) has pointed out in politics it is never enough even if the decision is "correct," it must be legitimate and accepted. Indeed, major policy advances will become possible only after public opinion, in this particular case that of teachers, local evaluation operators, and teachers' union leaders have been swayed to accept new ideas and the advantages and "benefits" of the evaluation, which due to its consequences, can clearly face powerful obstacles.

Majone (2005) argues that in the feasibility analysis of any policy it is not only necessary to make a calculation of optimal or better solutions within the given solutions, but also to discover instruments to expand the frontier of what is possible, which depends on what the political system and key actors in the policy process consider fair or acceptable. For this reason, he says, persuasion and conviction about the benefits of the policy that it seeks to promote are very important.

Those who exercise leadership in conducting the evaluation have the responsibility to generate the spaces to listen and communicate about the evaluation so that the actors can exercise their veto capacity, possible conflicts can be anticipated, and agreements and decisions generated that enable change to occur. If, due to poor communication, the discourse that teacher evaluation is punitive wins out, i.e., that it punishes and stigmatizes, it will face resistance from the most combative sections of the teachers' union, as happened in the Mexican case. Effective communication is required to face the dispute between the promises of teacher evaluation and its consequences. It is essential, then, for the implementation of a teacher evaluation policy to build a system for listening and communicating.

4.5 Final Considerations

In this last section, we conclude that the success of a teacher evaluation policy not only depends on its technical design, no matter how good it is, or even on an initial political agreement to carry it out, as happened in the Mexican case, but on how the areas of tension underlying the various stages of the policy cycle are resolved. This chapter emphasized that teacher evaluation policies do not always have the expected results, due to a series of tensions between the different actors involved in each stage of the policy cycle because of their different views on the purposes, types, and consequences of the evaluation, and also as a result of the timing during which different groups of actors expect the changes to occur.

Likewise, it was suggested that governance is required in the evaluation process, which takes into consideration the voices and participation of different social actors, including teachers and families. Additionally, effective communication regarding the benefits of the evaluation is a must. In this final section, we include a set of considerations, by way of conclusions, which we hope will contribute to a collective reflection on the importance of analyzing the politics of teacher evaluation policies.

4.5.1 The Evaluation is not Merely a Technical Exercise

The first consideration is to insist that educational evaluation in general, and in this case teacher evaluation, is not a merely technical exercise. It involves political considerations which are reflected in the tensions among all the actors who participate in its many levels. Ignoring this or not giving it due importance constitutes a blunder that will go against the very policy of improving teacher professionalization, and consequently its evaluation. Although the technical dimension of the evaluation must be unobjectionable, the negotiating capacity of those who act as leaders of the teacher evaluation policy(s) is irreplaceable and unavoidable in an agreed governance framework. For this, it is very important to have a map of actors, their possible gains, and losses, so that conflicts can be anticipated and negotiated.

4.5.2 It is Necessary to Clarify and Negotiate What the Teacher Evaluation is Meant to Accomplish

The educational system is a living, changing entity, and the teacher evaluation policy will necessarily have to be rooted in a clear, widely socialized and negotiated conception of what teacher professionalization means and its peculiarities in each society, and what the purpose of the teacher evaluation policy is. Although the salary dimension is unavoidable in any teacher evaluation system, there are other dimensions that must be considered and that are linked to the professionalization of the union, such as initial training and teacher professional development policies, which necessarily, should be linked to the evaluation process. It is argued that the essential component of a teacher evaluation policy should be its anchoring to the professionalization of teaching and the improvement of the teaching profession.

4.5.3 One of the Main Areas of Tensions Surrounding Teacher Evaluation Has to Do with the Different Visions and Interests of the Actors

The main tensions and conflicts present among the different actors involved in the evaluation are produced because of different understandings, views, or interests. Politicians tend to have a broad vision of the situation of an educational system with an emphasis on achieving specific goals with a teacher evaluation program, and in the face of the changing political scenario of forces and counterforces, it modulates their decisions and actions. The specialists, for their part, favor decisions about teacher evaluation made rigorously so that they are reliable and relevant in relation to the different components of the evaluation, from its design, through information gathering, its analysis and the presentation of global and individual results.

Sustaining policies aimed at improving the quality of education over time represents a complexity that is not easy to imagine. To this end, it is necessary to know the educational system, appreciate what is possible to achieve, accept what has not worked, and insist on the need for improvement with the implementation of one or another policy without forgetting the central purpose: the right of everyone to receive a quality education with equity. As has already been clearly stated, the issue is not whether there are contradictions and tensions, but how they are negotiated and resolved considering the greater good.

4.5.4 It is Necessary to Improve the Capacity of Listening and Governance of the Evaluation Process

Teacher evaluation will not achieve its objectives if there is not only a systematic and permanent process of listening, particularly to the teachers' union, since teachers are the object and direct recipients of the policy, but also the participation of different actors from civil society, who can play a counterweight role and demand the right to a quality education. Regarding teachers, as the consequences of the evaluation will affect them directly, it must be considered that they can be encouraged by union leaders or political operatives in the direction of opposing the evaluation or supporting it, and the mechanisms for one or another action to take place are varied and diverse. For this reason, as Corrales (1999) points out, "the inclusion of the possible beneficiaries of a policy in its design and implementation increases the probability of success of a policy" (p. 19).

Therefore, we must be open to the possibility of establishing new political agreements that can guarantee that the teacher evaluation policy is implemented, since as we have indicated, one thing is to design, and another is to make that design work in each reality. Listening to all the actors involved and getting them to participate in the process, in a permanent and systematic way, will make possible the required adjustments and new decisions.

4.5.5 We Must Provide Spaces and Mechanisms for Effective Communication About the Evaluation

As we have already pointed out, effective communication regarding the teacher evaluation and the benefits that this supposes for the field's professionalization and for the improvement of education overall is a necessary condition to ensure that it does not face greater resistance. Furthermore, in this sense, it is necessary to invest time and resources in communicating what is pursued with the evaluation, through different channels—digital media, printed material, etc.—but also, and importantly, with the different interest groups involved, perhaps starting with the establishment of advisory councils. Within these councils, all the available information can be presented, explaining the benefits of the evaluation that is to be implemented, discussing the direction it is taking and negotiating the agreements that are necessary to continue advancing.

Is it worthwhile to insist on undertaking the teacher evaluation process in the face of all the tensions and complexities mentioned? The answer is "yes" if the objective is the professionalization of teaching, and from there, the improvement of the educational system.

References

Aguilar, L. F. (2006). *Gobernanza y gestión pública.* México: Fondo de Cultura Económica.

Aguilar, L. F. (2010). *Introducción. Política pública.* México: siglo xxi editores. Available in http://data.evalua.cdmx.gob.mx/docs/estudios/i_pp_eap.pdf

Barber, M., & Mourshed, M. (2007). *How the world's best-performing school systems come out on top.* McKinsey & Company.

Barber, M., & Mourshed, M. (2008). *Cómo hicieron los sistemas educativos con mejor desempeño del mundo para alcanzar sus objetivos.* Santiago de Chile: PREAL.

Braslavsky, C., & Cosse, G. (1996). *Las actuales reformas educativas en América Latina: Cuatro actores, tres lógicas y ocho tensiones.* Santiago de Chile: PREAL.

Bruns, B., & Luque, J. (2014). *Profesores excelentes: Cómo mejorar el aprendizaje en América Latina y el Caribe.* Banco Mundial.

Corrales, J. (1999). *Aspectos políticos en la implementación de reformas educativas.* Santiago de Chile: PREAL.

Cuevas, Y., & Moreno, T. (2016). Políticas de evaluación docente de la OCDE: Un acercamiento a la experiencia en la educación básica mexicana. *Archivos Analíticos de Políticas Educativas*, 24, 1–20. Disponible en: https://www.redalyc.org/pdf/2750/275043450106.pdf

Easton, D. (1957). An approach to the analysis of political systems. *World Politics, 9*(3), 383–400.

Fernández, C. (2017). La reforma educativa en México, Chile, Ecuador y Uruguay. Aportes para un análisis comparado. *Revista Interamericana de Educación de Adultos*, 39(2), pp. 168–172.

Franco, H. (2019). Discursos sobre la evaluación del desempeño docente en el contexto de la reforma educativa de 2013 en México. *Edähi Boletín Científico de Ciencias Sociales y Humanidades del ICSHu, 8*(15), 37–49. Disponible en: https://repository.uaeh.edu.mx/revistas/index.php/icshu/article/view/4954/6852

Guzmán, C. (2005). Reformas educativas en América Latina: un análisis crítico. *Revista Iberoamericana De Educación, 36*(8), 1–12.

Hernández, O. D. (2021). *Los orígenes políticos de la reforma educativa 2013 en México.* Tesis para obtener el grado de Doctor en Ciencias Sociales y Políticas. Universidad Iberoamericana Ciudad de México: México.

INEE. (2016). *Evaluación del desempeño desde la experiencia de los docentes. Consulta con docentes que participaron en la primera Evaluación del desempeño 2015*. México: Autor.

INEE/OREALC/UNESCO. (2017). *Evaluación del desempeño de docentes, directivos y supervisores en educación básica y media superior de México. Análisis y evaluación de su implementación 2015–2016. Informe final.* México: INEE. Disponible en: https://www.inee.edu.mx/wp-content/uploads/2019/03/OREALC-UNESCO-Ev-desempeno-Informe-Final.pdf

Majone, G. (2005). *Evidencia, argumentación y persuasión en la formulación de políticas.* México: Fondo de Cultura Económica.

Manzi, J., González, R., & Sun, Y. (Eds.). (2011). *La Evaluación Docente en Chile.* MIDE UC, Centro de Medición de la Pontificia Universidad Católica de Chile: Santiago de Chile. Disponible en: https://www.mideuc.cl/web19/wp-content/uploads/Libro-Ev-Docente-en-Chile-FINAL-2011-07-20.pdf

Martinic, S. (2001). Conflictos políticos e interacciones comunicativas en las reformas educativas en América Latina, *Revista Iberoamericana de Educación. Monográfico Reformas Educativas: Mitos y Realidades/reformas Educativas: Mitos y Realidades, 27*, 17–33.

Martínez, F. (2016). *La evaluación de docentes de educación básica. Una revisión de la experiencia internacional*. INEE.

Merino, M. (2010). La importancia de la ética en el análisis de las políticas públicas, en M. Merino, y G. Cejudo (Eds.). *Problemas, decisiones y soluciones. Enfoques de políticas públic*as. México: CIDE/FCE.

Nava, M., & Rueda, M. (2014). La evaluación docente en la agenda pública. *Revista Electrónica de Investigación Educativa, 16*(1), 1–11. Disponible en http://redie.uabc.mx/vol16no1/contenido-nava-rueda.html

OCDE. (2010). *Mejorar las escuelas. Estrategias para la acción en México.* México: OECD Publishing.

Ravela, P. (2012). "La evaluación del desempeño docente para el desarrollo de las competencias profesionales" en Martín & Martínez (coord). *Avances y desafíos en la evaluación educativa.* Madrid: Fundación Santillana.

Rivas, A. (coord.) (2020). *Las llaves de la educación. Estudio comparado sobre la mejora de los sistemas educativos subnacionales en América Latina.* Madrid: Fundación Santillana.

Sabatier, P. A. (1988). An advocacy coalition framework of policy change and the role of policy-oriented learning therein. *Policy Sciences, 21*(2–3), 129–168.

Santibañez, L., Martinez, J. F., Datar, A., McEwan, P. J., Messan Setodji, C. & Basurto-Davila, R. (2007). *Haciendo camino: Análisis del sistema de evaluación y del impacto del programa de estímulos docentes Carrera Magisterial en México.* Secretaría de Educación Pública: México. Disponible en https://www.rand.org/pubs/monographs/MG471z1.html.

Schulmeyer, A. (2002). *Estado actual de la evaluación docente en trece países de América Latina.* Trabajo presentado en la Conferencia de los Maestros en América Latina y El Caribe: Nuevas prioridades. Brasilia, Brasil.

Sierra, L. (2017). *Análisis de la implementación de la evaluación del desempeño docente 2015 desde el enfoque de redes de política pública.* Tesis para obtener el grado de Maestra en Políticas Públicas Comparadas. Facultad Latinoamericana de Ciencias Sociales, Sede México.

Silva, S. (2017). Identificando a los protagonistas: el mapeo de actores como herramienta para el diseño y análisis de políticas públicas. *Gobernar: The Journal of Latin American Public Policy and Governance, 1*(1), 66–83.

Weiss, C. (2016). La investigación de políticas: ¿datos, ideas y argumentos?, en Banco de Desarrollo de América Latina (CAF). *La evaluación de políticas. Fundamentos conceptuales y analíticos. Serie Estado, Gestión Pública y Desarrollo en América Latina.* Buenos Aires, Argentina: CAF.

Chapter 5
Teacher Professionalism and Performance Appraisal: A Critical Discussion

Beatrice Ávalos

Abstract The chapter covers the relationship between concepts of teaching as a professional activity and approaches to teacher performance appraisal. In its first part, the chapter considers perspectives that cross discussions about teacher professionalism. It contrasts performative views of teaching (Ball SJ, J Educ Pol 18(2):215–228, 2003) and new public management policies with views of teachers as knowledge and practical professionals. These two approaches are expressed as differences between organizational and occupational professionalism (Evetts J, Current Sociol Rev 61(5–6:778–796, 2013). From an international perspective, the chapter deals with challenges to teachers' occupational professionalism in different contexts and examines research about this. More specifically, the chapter moves on to teacher evaluation developments in some national contexts and considers whether these mainly base their assessment criteria on teacher professionalism (formative) or on test-based learning outcomes (summative). The inclusion of teacher evaluation as part of formal career systems is discussed using (Tournier et al, Teaching career reforms: learning from experience, International Institute for Educational Planning, 2019)'s analysis of such systems, as well as studies that examine how teachers in different national contexts view their appraisal requirements. It concludes with a rephrasing of the notion of accountability that underlies teacher evaluation, in order to reclaim its meaning as a professional responsibility that teachers owe to those who respect and place trust in their work.

5.1 Introduction

This chapter has as its focus both the concept of teachers as professionals in the current policy contexts and how this professional character is or not upheld by approaches to teacher appraisal. It draws on sources in different world contexts that center on academic analysis of teacher policies as well as on studies dealing with teacher

B. Ávalos (✉)
Centre for Advanced Studies in Education (CIAE), University of Chile, Santiago, Chile
e-mail: bavalos254@gmail.com

J. Manzi et al. (eds.), *Teacher Evaluation Around the World*,
Teacher Education, Learning Innovation and Accountability,
https://doi.org/10.1007/978-3-031-13639-9_5

perceptions of the systems to which they are subject. This international focus is considered justified given the form in which teacher-related policies have travelled as have also related practices anchored on new public management and neo-liberal market policies. Specifically, besides examining longstanding analysis of teacher professionalism, the chapter is based on a literature review of recent studies on teacher professionalism and evaluation covering mainly, but not exclusively, from 2015 onward. While most studies occur in Anglophone countries or are published in English, an effort was made to include studies published in Spanish. Other limitations have to do with not having a wider international coverage with studies in Africa and Asia.

In discussing the notion of teachers as professionals and of teacher professionalism, the assumption is that teachers, by nature of their preparation and the complexity of their task, reassemble in their teaching sites their knowledge base—a mix of theory and practice—through analytic and reflective judgment about what students, as individuals and group, require to learn and do. The notion of "occupational professionalism" developed by Evetts (2013) aptly serves to describe this complex task. Further to this, the chapter takes on a discussion of challenges to teacher professional work derived from needing to guard their professionalism, support the quality of its enactment, and respond to what society expects from their teachers. For the task of education, teaching is a social obligation, as it is to ensure that every student has the opportunity to learn and develop. From this angle, the chapter discusses how appraisal or evaluation of teacher performance is researched, examines the procedures that support or narrow the scope of teacher responsibility to student test results, and how teachers respond to difficulties and sometimes threats to their professional occupation. In its concluding section, the article seeks to rephrase the concept of accountability as used to justify why teachers should be evaluated, in order to reclaim its meaning as a professional responsibility owed to those who trust their work.

5.2 Teaching—A Professional Occupation

Discussions centered on the nature of teaching have for long attempted to assert its status beyond earlier descriptions as being a quasi-professional activity (Hoyle, 1974; Etzioni, 1969). More recent studies on the nature of professional work have facilitated this analysis (Abbott, 1988; Freidson, 1989; Evetts, 2013), allowing teaching to be properly described as a professional occupation. Teachers can thus be referred to as professionals with a specific sphere of action defined as education and teaching, appropriate preparation, a related specific identity and a code of ethics. Teachers engage in work activities, rely on social recognition and trust, and exercise judgment based on appropriate knowledge and practical capacity (Abbott, 1988; McBeth, 2012; Swan et al., 2010; Yinger, 2005). As in other professional activities, what matters in the case of teachers is the legitimacy and quality of what they do, that is, their professionalism (Demirkasimoglu, 2010; Evetts, 2013; Goodson, 2003).

Teacher professionalism requires not only specific capacity for the job but also work toward its improvement. As with other professional occupations, beyond somewhat abstract definitions, a contested issue is the conditions under which teacher professionalism is monitored and protected: from "within" the occupational group or from "above", that is, by their educational systems' managers (Evetts, 2013). As shall be discussed later, this distinction is key in assessing the impact on teachers of New Public Management (NPM) and market-driven teacher policies (Hargreaves, 2000; Tolofari, 2005).

There are different views about what teacher professionalism entails in practice, how it develops through teacher education, and how it is enacted and protected in work situations (Demirkasimoglu, 2010). For example, while teaching is the field of action where teacher professionalism is at play, preparation for teaching may either accentuate its theoretical basis or on the contrary lay emphasis on its reflective pedagogic and practical elements, as illustrated by two contrasting teacher education programs in Germany studied by Dodilet et al. (2019). Teacher professionalism can also be viewed in relation to the historical evolution of teaching and of its tools and practices, as well as on how individual and collective teacher responsibility have played in its strategies and results. Along this process, teachers have engaged in transformative and collaborative forms of professionalism (Hargreaves, 2000; Hargreaves & O'Connor, 2017; Sachs, 2004). Achieved professional status, however, does not always entail professionalism in action (Ozga, 2000) as particular sociohistorical conditions may act as restrictive and/or as facilitating factors. To use a contemporary example, the abrupt change in the form of schooling and teaching brought about by the COVID-19 pandemic had two effects on teachers and their "lived" professionalism. The initial one, for many teachers around the world, can be described as an off-putting experience at the least and as a distressing one at its worst. What has followed, however, is an effort among teachers to collaboratively rework how they teach, utilizing instruments and approaches new to them in order to further their students' learning. These efforts can be aptly described as transformative and even creative expressions of teacher professionalism (Kim & Asbury, 2020; Niemi & Kousa, 2020).

Meanings of professionalism, how it is enacted and what level of control teachers have over its practical definition and monitoring, have evolved as referred to above. For example, Hargreaves (2000) wrote about a sort of "golden age" from the 1960s to the 1980s, mainly in Canada, the United States (USA), and the United Kingdom (UK), when teachers' working conditions supported "autonomous" and "collaborative" professionalism anchored on teacher continuous education. During this time, teachers were allowed a degree of freedom to implement curricula based on trust in their pedagogical competence to handle the demands of classroom teaching. Teachers were able to exhibit what Evetts' (2013) describes as "occupational professionalism", that is, professionalism defined and constructed by teachers and their profession. These conditions, however, were only partly operant in other world locations such as Latin America and Africa where teaching remained a non-graduate activity until well into the 2000s decade. Even where education conditions provided some space for teachers to exert professionalism, such as broad curricular frames and constructivist teaching approaches, as in Chile, Mexico, and South Africa in the early 2000s, teachers found it difficult to make use of these enabling contexts. This is due to limiting systemic conditions such as long teaching hours, narrow accountability pressures, and overcrowded classrooms (Ávalos, 2002).

5.2.1 Recent Challenges to Teacher Professionalism

With exceptions, it is difficult to signal out locations with "perfect" conditions that support teachers' work as professionals, that reward their work with just salaries, and provide sufficient leeway for them to respond to education needs as best as their preparation allows for. However, the emergence and spread of neo-liberal market and new public management (NPM) policies over world political systems have created conditions in the administration of public services affecting the work of teachers associated with them (Anderson, 2017; Ferlie, 2017). These policies have contributed to alter the understanding of education as a public good and foster the view that education services profit from being regulated by market forces. Specifically, regarding teachers, NPM policies advocate control over their competence based more on specifics of performance or "performativity" (Ball, 2003), rather than on a broad understanding of what is involved in teaching. Such policies support the monitoring of teacher performance with emphasis on accountability and standards, flexibility of teacher employment, and use of performance-based pay. In systems, as in Chile, where school funding is subject to student numbers, teachers as professionals find themselves conflicted in how best to handle their work as educators while responding to the external pressure of student examination results (Tolofari, 2005). In NPM contexts, teachers' voice and needs tend not to be sufficiently addressed, being regarded as objects of intervention rather than as subjects of change and feeling disempowered before families as the state takes over their broad decision-making power (Novaes & Silva, 2020; Van der Tuin & Verger, 2013).

New public management policies have not equally affected education systems. Most such policies originated and developed in Anglophone countries, mainly England, the USA, and New Zealand, but in the context of globalization (Rizvi & Lingard, 2010), these policies have influenced other locations with the market, neo-liberal political, and economic systems needed to sustain them, as is the case of Chile (Bellei & Vanni, 2015). Two recent studies that examine the geography of teacher-related policies illustrate how broad political and economic structures affect conditions for teacher professionalism. The first of these, based on teacher responses to the TALIS 2013 survey (Voisin & Dumay, 2020), reviewed models of teacher regulation covering initial education provisions, labor market structures, and division of labor. The resulting models and countries which fit these categories were classified in four groups that roughly represent the organizational and occupational professionalism types defined by Evetts (2013). Mainly professional models were identified in countries, such as Finland, Denmark, and Norway that place high value on teachers' professional knowledge and preparation as well as professional autonomy based on expertise. Market models accentuating standards-based regulation, diversity of teacher education pathways, as well as performance, managerial accountability, and low levels of teacher autonomy located in England, the USA, and Chile. The second study by Aoki and Rawat (2020) examined the extent of teacher performance pay, advocated by NPM policies, in 51 countries using questionnaire responses to the 2012 PISA study. Among, other characteristics, the authors distinguished between more or less "liberal" countries in political terms (i.e., stronger versus less strong democracies) and were able to show that performance-based pay tended to be used in less liberal systems, such as Singapore, Jordan, Thailand, and the Slovak Republic. Despite the origin of NPM policies in more liberal countries such as the USA, England, Australia, and New Zealand, performance-based pay has not been used there as much as the case might have been. The main thrust of NPM policies on teacher professionalism, particularly in England, the USA, and Chile, has derived from test-based school sorting and public funding that follows student numbers (Tolofari, 2005).

The 1988 Education Reform Act in the UK, which modified the school funding system on the basis of weighted per capita, sets the course for policies that impacted on education and teacher professionalism (Gewirtz et al., 1995). The later introduction of school accountability and rankings as well as the use of contextualized value-added measures (VAM) put pressure on teachers to secure a good positioning for their schools on league tables (Acqua, 2013). This policy environment practically obligated teachers to concentrate on the core subjects examined and to engage in teaching-to-the-test practices, thus lessening their professional discretion (Keating, 2015; Pring et al. in Acquah, 2013). In the USA education system, teacher evaluation based on generic performance criteria or standards was established following the A Nation at Risk policy (National Commission on Excellence in Education, 1983). Rationale for the system was a broad view of teacher professionalism (Danielson, 2007; Hunter, 1982). However, this approach to teacher evaluation was narrowed with the Federal Race to the Top initiative (RTTT, 2009). This policy introduced both value-added measures of teacher performance based on schools' test results and a narrower standards system (Danielson, 2016). Since 2015, the system has become less stringent in its accountability focus, as the different states are free to decide on how they evaluate their teachers (ESSA, 2015).

De-professionalizing NPM policies have had an effect in Australia (Sachs, 2004), Sweden (Hult & Edström, 2016) and selected locations in Asia, Africa, and Latin America (Kapucu, 2006). However, in some of these locations, information technology is altering the classical NPM form of public sector management producing a move toward what may be described as a bi-directional digital era of governance (Dunleavy et al., 2006). This change, which has become more noticeable with the impact of the global COVID-19 pandemic, offering new possibilities for teachers to respond professionally as individuals and collaboratively to what government managers require from them. Such responses may include professional interpretations of policy in line with what the teaching contexts require from them. A study of Australian teachers' response to demands posed by a new Literacy and Numeracy school testing system (Hardy et al., 2019) provides an example of such policy interpretation. The study focused on teachers who endeavored to assert their professionalism regarding the testing system's focus on data for its own sake and the short-term cycles expected for them to improve student results. They did so by denouncing the accountability system as diminishing their own professional capacity while also working more closely with students in need of attention. In other words, teachers responded to the policy by engaging in "intelligent" or "rich accountability" (Hardy et al., 2019). An example, also, of intelligent resistance to narrow accountability policies surfaced in an interview/questionnaire study with Swedish teachers (Hult & Edström, 2016). Teachers were asked how they perceived the effects of performance evaluations (international, national, and collegial/personal) and the accountability expectations these entailed. Contrary to what might be assumed, these teachers gave low ratings to the impact of such evaluations over their practice and were especially critical about external evaluations that reduced the possibility of being creative in their work. But on the other hand, teachers provided high ratings for their own school assessment results as providing food for reflective assessment about their practices, conducted on their own, with colleagues and with school principals.

Policy and decision-making in Canadian provinces and its education boards have been less influenced by NPM policies, although large-scale assessment is in place all over the country, and education authorities may link results to a diversity of teacher incentives. In this respect, a large survey and interview study by Copp (2017) brought out an effect of large-scale assessment over teachers' teaching to the curriculum and to the test. From a different perspective, Hardy and Melville (2019) conducted an interview study with educator members of the Ontario School Board in Canada on their understanding of teacher professionalism and their role regarding school policy. Throughout the interviews, a tension was observed as participants explained their criteria for assessing teachers' role in implementing a literacy and numeracy policy. This tension reflected competing forms of dealing with issues and demands of the policy, closer to organizational or to occupational forms of professionalism (Evetts, 2013). Thus, one group referred to criteria based on accountability, standardization of work, and student results in literacy and numeracy tests, that is, an organizational view of professionalism. On the other hand, the second group's opinions were closer to favoring teachers' autonomy, collegial authority, and professional ethics, that is, occupational professionalism.

These tensions between views that value teacher occupational professionalism, allowing for well-founded decision-making in teaching and school activities, and views that support organizational professionalism and the role of incentives associated to large-scale assessment results, mark much of the debate about the purposes and forms of teacher evaluation.

5.3 Teacher Performance Evaluation and Career Systems

Appraisal of teachers' work to verify its quality and assist in its improvement has for long been the task of school authorities or external inspectors and remains so in many countries. Interviews and direct observation of teaching also are the main instruments used for appraisal purposes. In its early forms, observation systems were simple in what they assessed and tended to approximate checklists of appropriate behaviors rather than respond to coherent views of teaching (Danielson & McGreal, 2000). However, toward the twenty-first century, conceptual work on teaching (Danielson, 1996; Eraut, 1994; Hunter, 1982; Marzano, 2007; Marzano & Toth, 2013)) helped to broaden the concept and assessment of teacher performance, thereby influencing evaluation systems toward establishing more comprehensive systems (Ávalos-Bevan, 2018; Clinton et al., 2016). Among the broad criteria frameworks used for evaluation purposes (Clinton et al., 2016) are adaptations of the Marzano Teacher Evaluation Model (Marzano & Toth, 2013), the Framework for Teaching Evaluation instrument (Danielson, 2011), and the Classroom Assessment Scoring System—CLASS (Pianta et al., 2008). The most common instruments for appraising teachers and providing them with feedback include teaching observations and portfolio evidence, although some systems also use student learning results provided by school or standardized tests.

Overtime, both evaluation policy and systems have been crossed by tensions arising from the extent to which they further occupational or organizational forms of teacher professionalism (Evetts, 2013). Thus, evaluation systems may have either mainly formative or accountability purposes and be associated with promotion and career stage allocation as well as demotion or dismissals (Tournier et al., 2019). Teacher evaluation policy in the USA exemplifies some of these tensions as do also teacher career system in various world locations.

The USA early formal teacher evaluation procedures derived from the A Nation at Risk Report (1983) largely rested on broad and generic descriptions of competent teaching performance such as provided by Danielson's (2007) framework. Based on generic descriptors and criteria, teacher assessment could include quality of lesson planning, of care for a classroom environment conducive to learning, of teaching strategies and how these responded both to curriculum orientations as well as students' differences, and finally on how they enacted professional responsibilities related to the school's community and relationships with parents. The introduction of teacher portfolios based on their work products also served to uphold teachers' professional role (Millman & Darling-Hammond, 1990). However, later modifications associated with the *No Child Left Behind Act* (2001) moved the focus of the evaluation system from teaching quality and professional responsibilities to student standardized test results expressed as value-added measures (VAM). Its negative effects on teacher professionalism and erosion of professional responsibility have been widely observed (Close et al., 2020; Jewell, 2017; Smith & Kubacka, 2017) including its effect over teaching to the test practices (Copp, 2017; Mintrop & Sunderman, 2013). The later *Every Student Succeeds Act* (ESSA, 2015) contributed to ease this focus on test results, leaving it to the different states to enact their own teacher evaluation systems.

Within this changing policy environment in the USA, there also are innovative deviations from narrow approaches to teacher evaluation that merit analysis. A comprehensive school-based approach to teacher performance evaluation not based on student results in the state of Cincinnati was examined in a school study that also observed its long-term effects over student learning (Taylor & Tyler, 2012). All teachers were evaluated every four years over one school year. During this time, teachers were observed three times by one of their peers and a fourth time by a school authority, receiving written feedback each time. Assessment of their work using Danielson's (2011) performance criteria included a summative score at the end of the year covering the four domains of the framework: preparation, classroom environment, teaching, and professional involvement in school and with parents. Teachers needing improvement were provided relevant assistance. To verify effects of the evaluation over student learning, Taylor and Tyler (2012) examined how teachers had impacted on their students' learning over two assessment periods, that is, ten years. Their results brought out a positive effect over student learning immediately after the evaluation year as well as in the following years, thus validating the effect of a well-thought-out form of evaluating teachers based on belief in their professionalism.

The extent to which systems of evaluation in other countries are enacted to further teacher professionalism varies. Over 90% of teachers participating in the TALIS 2013 survey reported that their schools' teacher evaluation included classroom observations as well as evidence from student tests, while a smaller number required evidence of content knowledge (Smith & Kubacka, 2017). In the later TALIS 2018 survey (OECD, 2020), 70% of teachers worked in schools that provided feedback about their performance based on student results (school/classroom) and/or students' external test results (65%). In many systems, head teachers are solely responsible for the appraisal of teachers, although in New Zealand, peers are also part of the teacher assessment system (Perry & Johns, 2018). In Finland, a very different system is in place and is of a clearly participatory and reflective nature (Woo, 2019). Teachers themselves conduct the process in line with their own development plan. School principals interact with teachers, discuss their plan, and support their professional development needs, all with a view of the coming school year rather than the past one. Consultations, of a participatory and reflective nature, also take place with peers.

In relation to systems of teacher performance evaluation, an OECD review in 18 countries (OECD, 2013) brought out a tendency to use performance evaluation with the purpose of holding teachers accountable to stakeholders more than as having formative goals. The review highlighted challenges such as the lack of a shared understanding of what is involved in high-quality teaching and use of appropriate evaluation procedures. Arguably, the report also suggested that country evaluation systems needed to find ways of considering student results in teacher appraisal and of using results to shape incentives for teachers (OECD, 2013). Among ways of addressing the challenges, the report recommended the consolidation of regular teacher developmental appraisal at school level, career-progression appraisal using external evaluators, standards to guide appraisal, and links with advancement decisions (OECD, 2013). While these recommendations might soften the impact of accountability-based evaluation, they do not remove the threats to teacher professionalism and mutual cooperation brought about by the association of performance evaluation to rewards and punishment measures.

5.3.1 Teacher Career Systems

Besides school-based teacher evaluation, different country systems have associated appraisal procedures with formal career progression stages thus potentially recognizing professional growth as well as teaching diversity. A study by Tournier et al. (2019) examined in ten countries a set of second-generation teacher career systems developed from the early 2000s onward in Colombia, Mexico, and Perú, as well as in Singapore, South Korea, South Africa, Thailand, Scotland, and the state of New York. To a large extent, these systems were influenced by NPM approaches and neo-liberal principles and include performance evaluation, ladders, and merit pay. Analysis of the ten systems as well as in-depth cases studies of three of them allowed the authors to highlight a diversity of issues related to their structure and enactment, while singling out the Scottish system as competent and well supported by teachers. Among recommendations for improvement, Tournier et al. (2019) included the need for clarity in the description of the evaluation criteria used, improvement of wording, and complexity in descriptions of profiles, parameters, and indicators, which seemed not to be the case in the South African and Mexican systems. Also problematic in some of the evaluation systems reviewed was the kind and number of the evaluation tools used. Thus, appropriate practices of classroom observation and interviews contrasted with dubious use of knowledge tests with multiple-choice items that were also highly criticized by teachers. Overall, according to the authors of the review, there is need for a good balance between the accountability and support purposes of teacher evaluation systems (Tournier et al., 2019).

One of Latin America's early systems was developed in Mexico in 1993 experiencing several changes since then (Guzmán, 2018). Initially, it established a voluntary five-level teachers' career together with a system of appraisal that would later include pay incentives. While maintaining the career system, legislation in 2013 made its evaluation compulsory for all teachers with results impacting on salaries and charged the newly created National Institute of Education (INEE) with conducting the process. An external evaluation of the system (Santiago, 2016) found it to be predominantly centered on accountability purposes rather than formative ones, with little attention given to teachers' work in the classroom and with limited participation of school authorities in the appraisal. Evaluator capacity also seemed insufficient. Changes in governmental policy since 2019 appear to diminish the accountability focus of the Mexican teacher evaluation by returning to the earlier more professional forms of career advancement (Santana, 2019; for more information, see Chapter Schmelkes in this same volume).

A more complex 5-stage teacher career and evaluation system is in place in Chile, regulated by legislation passed in 2016 (Ávalos-Bevan, 2018). The system combines professional development with accountability purposes. Progress through the career system, which includes salary increases at each stage, requires teachers to pass one test on school curriculum knowledge and to submit specified portfolio evidence for advance through all stages of the career. The first three career stages are compulsory ones. Failure to pass the evaluation after two tries is a cause for dismissal (for more information, see Sun chapter in this same volume).

5.3.2 Teacher Perception of Performance Appraisal Systems

The TALIS 2018 survey covering 48 countries (OECD, 2020) questioned teachers on the quality of feedback received from their appraisal experiences and how it affected their self-efficacy perceptions. Being appraised by more than one evaluator was related to teachers holding positive self-efficacy perceptions (in 23 countries). Feedback related to student test scores was associated with positive teacher self-efficacy (24 countries) as well as with job satisfaction (17 countries). Receiving feedback on classroom management affected self-efficacy in 17 countries and job satisfaction in 23 countries. On the other hand, feedback perceived as a mere administrative exercise was associated with lower teacher self-efficacy in 14 countries and lower job satisfaction in all participating countries.

Perhaps, the most contentious element of evaluations is their performative and less professional aspects, and the degree to which the system is high stakes and impacts on teachers' stress and well-being. In this respect, a survey of 1.866 teachers in three USA states (Ryan et al., 2017) found that the accountability systems in use in two of the states and planned for the third one, significantly predicted situations of stress, burnout, and intention to leave the profession on the part of teachers. A similar situation of discomfort was brought out by teachers in the state of Río de Janeiro in Brazil, where school and teacher evaluation established between 2009 and 2014 used VAM scores. Interviews with teachers brought out their apprehensions about having to set aside what they termed as a pedagogic approach to learning in order to respond to the VAM's emphasis on test results: "with all this pressure we stop thinking of students as students, as people with individual needs and concerns. They become metrics to be increased". (Straubhaar, 2017, p. 12)

In Sweden, where teachers are subject to several forms of evaluation, an interview study with 34 teachers from municipal and independent schools recorded their diverse concerns about the system (Hult & Edström, 2016). Compared to school evaluations performed by teachers, those interviewed found that external ones were less pertinent and time-consuming. In their view, these assessments do not allow them to be as creative and independent in their work as do school-based ones and felt that the system was based on mistrust about their capacity. As concluded by Hult and Edström (2016), the interviews reflected a clash between teacher professional responsibility and the external accountability demands to which teachers felt subjected. In Chile, a similar interview study with 60 primary and secondary public school teachers provided evidence of tensions between their professional identities and having to submit to external evaluation of their work (Sisto, 2011). Teachers believed that those who judged their performance lacked inside or relevant knowledge about their teaching and school circumstances, however, "expert" or "knowledgeable" they might be. On the other hand, the teachers interviewed appreciated the relevance of school appraisals for being conducted by authorities who not only know the school but also value effective forms of teaching and learning. As concluded by Sisto (2011), the Chilean external teacher evaluation clashes somehow both with teachers' historical identity as collaborative professionals and a developing new identity, as responsible and accountable professionals within their school community. In other words, the teachers studied did accept the need for performance appraisal, but as a school embedded process and not as an externally conducted one.

Another study in Chile (Acuña, 2015) explored teacher views regarding the content knowledge test which was part of the evaluation system until its changes in 2016 and taken voluntarily by those aspiring to a pay incentive for successful performance. By means of focus groups and interviews, the study inquired how teachers perceived this appraisal system and how much sense it made to them to be eligible for economic incentives associated with good performance. Arising from the data, Acuña (2015) distinguished four types in how teachers associated monetary incentives with their perceived roles. The first type was teachers who valued as such the social role of teaching regardless of its possible impact on salary bonuses. The second type identified themselves as part of a knowledge-based profession insufficiently rewarded by their salary scheme and therefore felt bonuses were justified. The third group were "saviors" who saw their role as helping students cope, face, and overcome their liabilities. These teachers did not expect incentives for their work. The fourth type represented professionalism in action, being teachers who were moved by student values' development, learning, or both and deserved an appropriate salary. However, as a group, these teachers questioned the notion of measuring and rewarding their work with monetary incentives. Though did not object to these incentives, these were accepted as a low-level substitute for a just salary that as professionals they should and were not receiving (Acuña, 2015).

5.4 Reflections and Conclusions

An important purpose of this chapter was to bring out and support the notion of school teaching as a professional occupation and of teachers as professionals (Evetts, 2013) in the context of policies associated with performance evaluation. Embracing this position might appear as a repetitive return to arguments over fifty years ago based on definitions of teaching as a "quasi-profession" (Etzioni, 1969) and more recently as a professional activity or occupation (Evetts, 2014; Yinger, 2005). The discussion, however, is valid and evident in current education policy analysis. The global impact of new public management and neo-liberal market policies have rekindled concerns about teacher professional work and its extent and limits (Anderson, 2017). The "occupational" professionalism of teachers as conceptualized by Evetts (2016) appears contested when claims for ownership and monitoring of teacher work are narrowed to externally measured results (Smith & Kubacka, 2017).

In relation to the above threats, the concept of teacher professionalism is benefiting from recent and more sophisticated analysis that describes teachers as knowledge workers (Price & Weatherby, 2021), affirming their key traits vis-à-vis restrictive views of what is expected of them. The quality of teachers' work rests on a knowledge base acquired through solid initial preparation and broadened through a variety of professional development activities. This knowledge gives form both to the teaching of curriculum content and to the pedagogy that teachers use to reach and support students and their learning. Enactment of their knowledge base in practice is complex, more so at the beginning stages of a teacher's career. However, it is not a solitary task, but the joint task of teachers and their school community. Teachers assert this view of the profession when they object to evaluations that value only a limited range of what they do. As knowledge workers charged with a social task, teachers appreciate a wider social recognition of the scope of their work, which is also central to their well-being perceptions (Acuña, 2015).

Accountability is a term with negative connotations in teacher evaluation policy analysis. In part, this perception brings out the "datafication" implications of appraisal systems that reduce the wider scope of teaching activities. This is especially relevant with respect to VAM teacher evaluation. Yet, of itself, the concept of accountability need not be cast aside. To demystify the notion, the recent 2017/2018 Education Global Monitoring Report (UNESCO, 2017) adopted the concept of accountability as its main theme and broadened its meaning. The report describes accountability along three main elements: (a) clearly defined responsibilities; (b) obligation to provide an account of how such responsibilities are met; and (c) legal, political, social, or moral justification for the obligation to account (UNESCO, 2017, p. 4). Extending this concept to teaching as a professional occupation (Evetts, 2013) and to teachers as knowledge workers in schools and classrooms (Price & Weatherby, 2021), the rationale for teacher accountability claims should derive from their mission, their agreed-upon duties, and the legal system under which teachers work. As this chapter brought out, teachers can face threats to their professionalism by enacting "intelligent accountability" that upholds the broad social orientation of the education while individually and collaboratively monitoring the quality of their teaching (Hardy et al., 2019). And do so in schools with well-organized systems of teacher assessment and clear formative feedback (Taylor & Tyler, 2012).

Systems of teacher appraisal centered on how teachers conduct their work in situ validate teacher "accountability" both as an instrument for feedback and improvement, as well as information for career progression. But, narrowing the evidence and procedures by which teacher accountability is claimed attempts against teachers as responsible knowledge professionals. In that respect, rather than continue to use an arguable word, it might be useful to replace the notion of teacher evaluation as an accountability obligation with the concept of appraisal as a "professional responsibility" (Fenwick, 2016 in Anderson, 2017).

Following on the above, there are many education systems that avoid the most questioned forms of teacher evaluation which are based on narrow standards and student learning scores, while using strategies that further teacher professionalism (Clinton et al., 2016). These systems enact appraisal procedures that include both observation of teacher classroom teaching as well as selected evidence of their work that teachers themselves gather, as in portfolios. In these systems, the location of appraisal is mainly in the school and its conduction is a responsibility of school authorities and may involve teacher peers. These forms of appraisal are guided by systems of standards developed at national or state level that represent an expression of what teachers know and can do in their classrooms and schools. There are good examples of such procedures in different parts of the world. For example, based on a review of six country systems, Perry and Johns (2018) brought out the case of Singapore labelling it as highly sophisticated. While the teacher evaluation system is national and centralized, its foundation rests in the school. Teacher performance appraisal includes classroom observation by a school supervisor, portfolio self-evidence, peer consultation, and student results. Schools foster a strong collaborative culture thus moderating the concept of performance as mainly an individual's responsibility. Similar examples were included in Tournier et al. (2019) review of teacher career systems.

To conclude, it is difficult to reconcile those views of teacher professionalism discussed in the first part of this article, with accountability demands based on narrow performance appraisal that overlooks the complexity of teaching and inordinately associates student test results with teaching quality. However, responsible accountability as described in the GMR Report (UNESCO, 2017) suggests that teacher appraisal anchored on respect for teachers as knowledge professionals, on student learning as jointly influenced by the school teaching community and conducted where teaching takes place has the potential to improve the quality of teaching and the learning of students. As expressed by a noted English educator (Whitty, 2000) to move in this direction requires demystifying teacher professional work. It requires teaching to be more democratic in its construction and appraisal, with parents, students, and the community as participants, thus, counterbalancing the narrow accountability demands operating in the context of market competitiveness (Whitty, 2000).

Acknowledgements Support from PIA-CONICYT Basal Funds for Centers of Excellence Project FB0003 is gratefully acknowledged.

References

Abbott, A. D. (1988). *The system of professions: An essay on the division of expert labor*. The University of Chicago Press.

Acqua, S. (2013). *School accountability in England: Past, present and future.* Technical Report. Manchester: Centre for Education Research and Policy. https://www.researchgate.net/publication/323611355_School_Accountability_in_England_Past_Present_and_Future

Acuña Ruz, F. (2015). Incentivos al trabajo profesional docente y su relación con las políticas de evaluación e incentivo económico individual. *Estudios Pedagógicos XLI, 1*, 7–26.

Anderson, G. (2017). Privatizing subjectivities: How new public management (NPM) is designing a "new" professional in education, 33(3), 593–626. https://doi.org/10.21573/vol33n32017.79296

Aoki, N., & Rawat, S. (2020). Performance-based pay: Investigating its international prevalence in light of national contexts. *American Review of Public Administration, 50*(8), 865–879. https://doi.org/10.1177/0275074020919995

Ávalos-Bevan, B. (2018). Teacher evaluation in Chile: Highlights and complexities in 13 years of experience. *Teachers and Teaching: Theory and Practice, 24*(3), 397–311. https://doi.org/10.1080/13540602.2017.1388228

Ávalos, B. (2002). How do we do it? Global rhetoric and the realities of teaching and learning in the developing world. In C. Sugrue & C. Day (Eds.) *Developing teachers and teaching. International research perspectives.* Routledge.

Ball, S. J. (2003). The teacher's soul and the terrors of performativity. *Journal of Educational Policy, 18*(2), 215–228. https://doi.org/10.1080/0268093022000043065

Bellei, C., & Vanni, X. (2015). The evolution of educational policy, 1980–2014. In S. Schwartzman (Ed.), *Education in South America* (pp. 179–200). Bloomsbury.

Close, K., Amrein-Beardsley, A., & Collins. (2020). Putting teacher evaluation on the map: An overview of states' teacher evaluation systems post-every students succeeds Act. *Education Policy Analysis Archives, 28*(58). https://doi.org/10.14507/epaa.28.5252

Copp, D. T. (2017). Policy incentives in Canadian large-scale assessment: How policy levers influence teacher decisions about instructional change. *Education Policy Archives, 25*(115). https://doi.org/10.14507/epaa.25.3299

Danielson, C. (2007). *Enhancing professional practice: A framework for teaching* (2nd. ed.). Alexandria, Va: ASCD.

Danielson, C. (2011). Evaluations that help teachers learn. *Educational Leadership, 68*(4), 35–39.

Danielson, C. (2016, April 18). It's time to rethink teacher evaluation. *Education Week*.

Danielson, C., & McGreal, T. (2000). *Teacher evaluation to enhance professional practice.* Alexandria, Va: ASCD.

Demirkasimoglu, N. (2010). Defining "teacher professionalism" from different perspectives. *Procedia Social and Behavioral Sciences, 9*, 2047–2051. https://doi.org/10.1016/j.sbspro.2010.12.444

Dodilet, S., Lundin, S., & Krüger, J. O. (2019). Constructing professionalism in teacher education. *Education Inquiry, 10*(3), 208–225. https://doi.org/10.1080/20004508.2018.1529527

Dunleavy, P., Margetts, H., Bastow, S. & Tinkler, J. (2005). New public management is dead-long live digItal-era governance. *Journal of Public Administration Research and Theory, 16*(3), 467–494. https://doi.org/10.1093/jopart/mui057

ESSA. (2015). US Department of Education. Retrieved on January 7, 2021, from https://www.ed.gov/essa?src=rn

Etzioni, A. (Ed.). (1969). *The semi-professions and their organization.* Free Press.

Evetts, J. (2013). Professionalism: Value and ideology. *Current Sociology Review, 61*(5–6), 778–796. https://doi.org/10.1177/0011392113479316

Evetts, J., et al. (2014). The concept of professionalism: Professional work, professional practice and learning. In S. Billet (Ed.), *International handbook of research in professional and practice-based learning* (pp. 29–56). Springer.

Ferlie, E. (2017). The new public management and public management studies. *Oxford Research Encyclopedias*. https://doi.org/10.1093/acrefore/9780190224851.013.129

Gewirtz, S., Ball, S. J., & Rowe, R. (1995). *Markets, choice and equity in education.* Open University Press.

Goodson, I. F. (2003). *Professional knowledge, professional lives. Studies in education and change.* Open University Press.

Guzmán Marín, F. (2018). La experiencia de la evaluación docente en México. Análisis crítico de la imposición del servicio profesional docente. *Revista Iberoamericana de Evaluación Educativa*, 11(1), 135–158. https://doi.org/10.15366/riee2018.11.1.008

Hardy, I., & Melville, W. (2019). Professional learning as policy enactment: The primacy of professionalism. *Education Policy Archives, 27*(90). https://doi.org/10.1080/13540602.2017.1388228

Hardy, I., Reyes, V., & Hamid, M. O. (2019). Performative practices and 'authentic accountabilities': Targeting students, targeting learning? *The International Education Journal: Comparative Perspectives, 18*(1), 20–33. https://openjournals.library.sydney.edu.au/index.php/IEJ

Hargreaves, A. (2000). The four ages of professionalism and professional learning. *Teachers and Teaching, 6*(2), 151–182. https://doi.org/10.1080/713698714

Hargreaves, A., & O'Connor, M. T. (2017). *Collaborative professionalism.* WISE and Boston College.

Hult, A., & Edström, C. (2016). Teacher ambivalence toward school evaluation: Promoting and ruining teacher professionalism. *Education Inquiry,7*(3), 305–325. https://doi.org/10.3402/edui.v7.30200

Hunter, M. (1982). *Mastery teaching*. Corwin Press.

Jewell, J. W. (2017). From inspection, supervision and observation to value-added evaluation: A brief history of US teacher performance evaluation. *Drake Law Review, 65*, 363–419. https://lawreviewdrake.files.wordpress.com/2015/01/jewell-final.pdf

Kapucu, N. (2006). New public management: theory, ideology and practice. In A. Farzamand & J. Pinkowski (Eds.) *Handbook of globalization, governance, and public administration* (pp. 889–901). Routledge.

Keating, I. (2015). English case study. Professional autonomy, accountability and efficient leadership and the role of employers organisations, trade unions and school leaders. Retrieved on January 7, 2021, from https://www.csee-etuce.org/images/attachments/RP_Professional_Autonomy_Accountability.pdf

Kim, L. E., & Asbury, K. (2020). 'Like a rug had been pulled from under you': The impact of COVID-19 on teachers in England during the first six weeks of the UK lockdown. *British Journal of Educational Psychology, 90*, 1062–1083. https://doi.org/10.1111/bjep.12381

Marzano, R. J., & Toth, M. D. (2013). *Teacher evaluation that makes a difference: A new model for teacher growth and student achievement*. Alexandria, VA: ASCD.

McBeth, J. (2012). *Future of the teaching profession.* Educational International Research Institute and Faculty of Education, University of Cambridge.

Millman, J., & Darling-Hammond, L. (Eds.). (1990). *The new handbook of teacher evaluation. Assessing elementary and secondary school teachers.* Corwin Press Inc.

Mintrop, H., & Sunderman, G. L. (2013). The paradoxes of data-driven school reform. In D. Anagnostopoulos, S. A. Rutledge, & R. Jacobsen (Eds.), *The infrastructure of accountability. Data use and the transformation of American education.* Harvard Education Press.

National Commission on Excellence in Education. (1983). A nation at risk: The imperative for educational reform. *Elementary School Journal, 84*(2), 113–130.

Niemi, H. M., & Kousa, P. (2020). A case study of students' and teachers' perceptions in a Finnish high school during the COVID Pandemic. *International Journal of Technology in Education and Science (IJTES), 4*(4), 352–369.

No Child Left Behind. (2001). Elementary and Secondary Education Act. US Department of Education. https://www2.ed.gov/nclb/landing.jhtml

Novaes, L. C., & Silva. T. M. M. (2020). As recomendações de organismos internacionais na política educacional paulista. *Arquivos Analíticos de Políticas Educativas*, *28*(175). https://doi.org/10.14507/epaa.28.4389

OECD. (2013). *Synergies for better learning: An international perspective on evaluation and assessment*. OECD Publishing.

OECD. (2014). *TALIS 2013 results: An international perspective on teaching and learning*. OECD Publishing.

OECD. (2020). *TALIS 2018 results (Volume II): Teachers and school leaders as valued professionals*. OECD Publishing.

Ozga, J. (2000). *Policy research in educational settings. A contested terrain.* Open University Press.

Perry, T., & Johns, P. (2018, September). Evaluating English teacher evaluation. How does teacher evaluation policies in England compare to international policy, practice and evidence. In *Presentation British educational research association meeting*. http://www.curee.co.uk/files/publication/%5Bsite-timestamp%5D/Teacher%20Evaluation%20in%20England%20-%20BERA%202018_0.pdf

Pianta, R. C., La Paro, K. M., & Hamre, B. K. (2008). *Classroom assessment scoring system (CLASS) manual*. Paul H. Brookes Publishing Company.

Price, H. E., & Weatherby. (2021). The status of teachers: How does treating teachers as knowledge workers influence their perception of value? In *Paper presented at the comparative and international education society, 65th annual meeting, 2021: Social responsibility within changing contexts*, April 25–May 2.

Rizvi, F., & Lingard, B. (2010). *Globalizing education policy*. Routledge.

RTTT. (2009). US Department of Education. Retrieved on January 7, 2021, from https://www2.ed.gov/programs/racetothetop/index.html

Ryan von der, E., Pendergast, S., Segool, & Schwing. (2017). Leaving the teaching profession: The role of teacher stress and educational accountability policies on turnover intent. *Teaching and Teacher Education, 66*, 1–11. https://doi.org/10.1016/j.tate.2017.03.016

Sachs, J. (2004). *The activist teaching profession.* Open University Press.

Sachs, J. (2016). Teacher professionalism: Why are we still talking about it? *Teachers and Teaching: Theory and Practice, 22*(4), 413–425. https://doi.org/10.1080/13540602.2015.1082732

Santana, A. (2019, May 29). Reflexiones sobre los nuevos cambios a la evaluación docente. *Arena Pública.* https://www.arenapublica.com/blog-alicia-santana/reflexiones-sobre-los-nuevos-cambios-la-evaluacion-docente

Santiago, P. (2016). Políticas de evaluación educativa. Evaluación docente en Mexico: las recomendaciones de la OCDE. In G. Niebla, Meléndez Irigoyen, M. T., Ramón Castaño, F. E., H. Sánchez Pérez, & F. Tirado Segura (Eds.), *La Evaluación Docente en el Mundo.* Fondo Cultura Económica & INEE.

Sisto, V. (2011). Nuevo profesionalismo y profesores: una reflexión a partir del análisis de las actuales políticas de 'profesionalización' para la educación en Chile. *Signo y Pensamiento, 30*(59), 178–192. 10.11144/.

Smith, W. C., & Kubacka. (2017). Emphasis on student test scores in teacher appraisal systems. *Education Policy Analysis Archives, 25*(86). https://doi.org/10.14507/epaa.25.2889

Straubhaar, R. (2017). The "power" of value-added thinking: Exploring the implementation of high-stakes teacher accountability policies in Rio de Janeiro. *Education Policy Analysis Archives, 25*(91). https://doi.org/10.14507/epaa.25.3034

Swann, M., McIntyre, D., Pell, T., Hargreaves, L., & Cunningham, M. (2010). Teachers conceptions of teacher professionalism in England in 2003 and 2006. *British Educational Research Journal, 36*(5), 549–571. https://doi.org/10.1080/01411920903018083

Taylor, E. S., & Tyler, J. H. (2012). The effect of evaluation on teacher performance. *American Economic Review, 102*(7), 3628–3651. https://doi.org/10.1257/aer.102.7.3628

Tolofari, S. (2005). New public management and education. *Policy Futures in Education, 3*(1), 75–89.

Tournier, B., Chimier, C., Childress, D., & Raudonyte, I. (2019). *Teaching career reforms: Learning from experience*. International Institute for Educational Planning.

UNESCO. (2017). *Global monitoring report 2017/8: Accountability in education: Meeting our commitments*. UNESCO.

Van der Tuin, M., & Verger, A. (2013). Evaluating teachers in Peru: Policy shortfalls and political implications. In A. Verger, H. K. Altinyelken, & M. de Koning (Eds.), *Global managerial education reforms and teachers: Emerging policies, controversies and issues in developing contexts*. Education International.

Verger, A., Altinyelken, H. K., & de Koning, M. (Eds.). (2013). *Global managerial education reforms and teachers: Emerging policies, controversies and issues in developing contexts*. Education International.

Voisin, A., & Dumay, X. (2020). How do educational systems regulate the teaching profession and teachers' work? A typological approach to institutional foundations and models of regulation. *Teaching and Teacher Education*. https://doi.org/10.1016/j.tate.2020.103144

Whitty, G. (2000). Teacher professionalism in new times. *Journal of in-Service Education, 26*(2), 281–295. https://doi.org/10.1080/13674580000200121

Woo, H. (2019, December 16). Professional responsibility of teachers: Teacher evaluation in Finland. *Teacher Evaluation Policy Series. Forum of the American Journal of Education*. https://www.ajeforum.com/professional-responsibility-of-teachers-teacher-evaluation-in-finland-by-hansol-woo

Yinger, R. J. (2005). A public politics for a public profession. *Journal of Teacher Education, 56*(3), 2385–3290. https://doi.org/10.1177/0022487105275921

Part II
Teacher Evaluation Systems Around the World: North America

Chapter 6
Teacher Evaluation in Washington, DC Public Schools

Aliza Husain, Jessalynn James, and James Wyckoff

Abstract IMPACT, the Washington, DC Public Schools (DCPS) teacher evaluation system, grew out of a longstanding frustration with unacceptably low student outcomes, creating a political climate open to substantial reform. In 2007, the District of Columbia passed the Public Education Reform Amendment Act (PERAA) that transitioned DCPS to mayoral control and led to a slurry of reforms—most contentious of which was IMPACT. IMPACT's design was informed by research evidence and intended to improve teaching quality and student outcomes through several mechanisms. It includes multiple measures of teaching effectiveness, opportunities for teacher feedback and development, and meaningful consequences for very weak or very strong teacher performance. Evidence demonstrates that IMPACT has substantially improved the quality of teachers and teaching, and consequently students' academic achievement. Evaluation systems, such as IMPACT, involve trade-offs. Some teachers may leave their jobs due to the stress associated with high-stakes evaluation. Other teachers value the recognition and development opportunities that IMPACT brings. DCPS has revised IMPACT over time, in response to evidence as well as stakeholders' concerns. An essential question facing DCPS is how to continue to redesign elements of IMPACT to better support teaching and learning while maintaining the benefits it has brought to teachers and students.

A. Husain (✉)
Pivot Learning, Oakland, CA, USA
e-mail: ahusain@pivotlearning.org

J. James
TNTP, New York, NY, USA
e-mail: jessalynn.james@tntp.org

J. Wyckoff
University of Virginia, Charlottesville, VA, USA
e-mail: jhw4n@virginia.edu

J. Manzi et al. (eds.), *Teacher Evaluation Around the World*,
Teacher Education, Learning Innovation and Accountability,
https://doi.org/10.1007/978-3-031-13639-9_6

6.1 Background and Context

By 2007, the District of Columbia had experienced several decades of poor student academic performance (NRC, 2011). The backdrop to the city's struggling public education system was a constantly changing governance structure, a series of financial and management scandals, and blatant racial and economic segregation. Such disfunction led to an inequitable distribution of resources, harming those students who needed the most supports, as well as a growing charter school sector (NRC, 2011).

In response, the city council passed the Public Education Reform Amendment Act of 2007 (PERAA; NRC, 2011). At the time, the District of Columbia Public Schools (DCPS) was among the nation's lowest-performing districts. For example, DCPS had the lowest 4th-grade math score of the 11 districts participating in the 2007 National Assessment of Educational Progress's (NAEP) Trial Urban District Assessments. It also had the lowest 8th-grade math score, the second lowest 4th-grade reading score, and the second lowest 8th-grade reading score (USDOE, 2007). These scores illuminated a dire need for rapid and significant change, and PERAA was designed to address these issues.

PERAA was already the second reform of its nature since 2000, but arguably the most dramatic, bringing significant structural reforms to education in DCPS. PERAA redistributed control of the city's public schools and public charter schools from the city's elected school board to the mayor and created a new state department of education. Also created were the role of the Deputy Mayor of Education, the Office of the State Superintendent of Education (OSSE), the State Board of Education (SBOE), and the Public Charter School Board. Finally, the person appointed to the newly created chancellor role would report directly to the mayor (NRC, 2011).

Mayor Adrian Fenty, elected in 2007 on a platform that stressed education reform, became the first mayor with control over DCPS. Fenty appointed Michelle Rhee as the first chancellor of DCPS (NRC, 2011). Rhee, the CEO and founder of The New Teacher Project (TNTP), an organization that focused on recruiting talented teachers to high-need schools, had worked with DC school policymakers for several years and was acutely aware of the problems DCPS was facing (Toch, 2018). She assembled a leadership team that included Kaya Henderson as the Chief of Human Capital in DCPS—who later became Rhee's successor as Chancellor—and Jason Kamras, a former DCPS teacher, as Henderson's second in command. Both had extensive experience in DC. Specifically, Henderson had directed Teach for America's DC program and had worked with Rhee at TNTP, while Kamras had taught in DC for more than a decade and had recently been named the national Teacher of the Year (Toch, 2018).

Under the new leadership team, DCPS immediately implemented substantial changes, most of which were focused on the quality of the district's teacher workforce (NRC, 2011). To begin, several hundred teachers were dismissed (Toch, 2018) along with about 30 principals (20%) in 2007–08 for poor performance (NRC, 2015). DCPS also hired more than 100 instructional coaches at a cost of $13 million a year

(Toch, 2018). Finally, 15 low-enrollment schools were closed to more efficiently allocate the district's resources (NRC, 2015).

The theory of change guiding these reforms hypothesized that improving teaching quality would raise student outcomes. The new leadership believed that a rigorous teacher evaluation system was central to these efforts. Consequently, DCPS replaced the existing teacher evaluation system, which rated 95% of teachers as satisfactory or better, with IMPACT, a new system based on emerging research on effective teacher evaluation. IMPACT was premised on two mechanisms—differential retention of effective teachers and supports for teachers to improve. It included high-stake incentives for teachers judged to be low and high performing (Toch, 2018).

Because Congress had given control of DC's teacher evaluations to the school board, the district did not have to negotiate with the Washington Teachers Union (WTU) to implement the evaluative component of IMPACT. They did, however, have to get buy-in from WTU to implement pay-for-performance. With the collective bargaining contract expiring in 2007 and the mayor backing IMPACT, the WTU was already on weak footing entering conversations with Rhee. Nevertheless, several years of tense negotiations ensued, following which Rhee provided DCPS teachers with an average 22% pay raise over a period of five years; in exchange, the WTU agreed to IMPACT's implementation. IMPACT was officially rolled out in the 2009–10 school year and was implemented across DCPS without pilot testing given the urgency with which the need for reform was viewed, and concern that momentum would be lost if it were not rolled out to all schools at the same time (Toch, 2018).

6.2 Designing Impact

The design of IMPACT began in 2007, with DCPS spending about $1.5 million dollars on its initial development (Toch, 2018). The individuals charged with designing IMPACT, Kaya Henderson and Jason Kamras, studied the current research on teacher evaluation and teacher quality and traveled across the country to meet with experts on the subject (Toch, 2018), while also engaging in focus groups with their current teachers and school administrators (NRC, 2015). The system they developed incorporated cutting-edge research on teacher evaluation to address the immediate needs of DCPS students.

IMPACT was designed to affect three levers of teacher quality: recruitment, development, and retention (Fig. 6.1). Each of these mechanisms and ensuing design features were informed by what was at the time an emerging evidence base. DCPS recognized that teachers are pivotal for student success, but that urban schools with histories of poor academic achievement and high poverty are often passed over by the typical teacher in favor of higher-achieving, lower-poverty schools (Adnot et al., 2017; Boyd et al., 2010, 2013; Jackson, 2009, 2013). Schools with such profiles likewise struggle to retain effective teachers, leading to cycles of understaffing and high turnover that perpetuate low academic achievement (Guarino et al., 2011; Ronfeldt et al., 2013). Meanwhile, much of the professional development typically used by

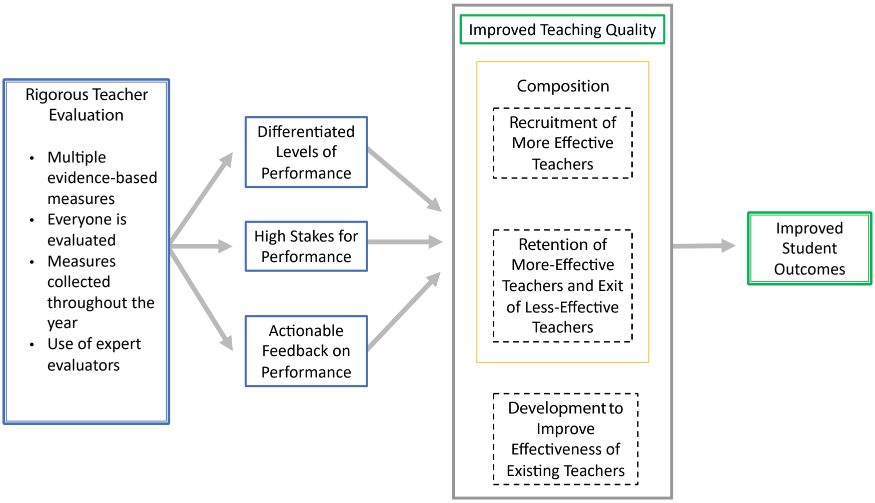

Rigorous Teacher Evaluation
• Multiple evidence-based measures
• Everyone is evaluated
• Measures collected throughout the year
• Use of expert evaluators
Differentiated Levels of Performance
High Stakes for Performance
Actionable Feedback on Performance
Improved Teaching Quality
Composition
Recruitment of More Effective Teachers
Retention of More-Effective Teachers and Exit of Less-Effective Teachers
Development to Improve Effectiveness of Existing Teachers
Improved Student Outcomes

Fig. 6.1 IMPACT theory of change

U.S. school districts at the time failed to demonstrate meaningful effects on teaching quality or student learning (Wayne et al., 2008; Yoon et al., 2007).

DCPS policymakers had good reason to believe that improved teaching had the potential to improve student outcomes. Before IMPACT, 95% of DCPS's teachers were rated as having met or exceeded expectations, yet the vast majority of its students were performing below grade level (Toch, 2018), and, as documented above, DCPS compared unfavorably to other districts on the NAEP. The district's abysmal academic achievement and extraordinarily low graduation rates lent a sense of urgency to the new reforms. A driving consideration for IMPACT's design was the recognition of the deep and widespread harm to children who were not being adequately served by their schools. The new system needed to better serve students by improving the quality of teachers, even if that meant dismissing some teachers judged to be underperforming. Critics were concerned that the system would misjudge effectiveness, and sanction teachers inappropriately. IMPACT attempted to minimize the risk of misidentifying teachers by employing several measures of teacher performance and by giving low-performing teachers supports and an opportunity to improve. As such, IMPACT aimed to balance a robust system of teacher evaluation and supports for teachers while holding teachers accountable. This balancing act continues to play into adaptations to IMPACT from 2009 to 2020.

For both political and substantive reasons, every DCPS employee is subject to a form of IMPACT evaluation. We focus on teacher evaluation, as this has received the most attention and is the primary mechanism by which policymakers envisioned improvement. By differentiating teachers' performance, DCPS could provide teachers with appropriate supports (if low performing) or recognition of excellence (if high performing). Chronically or exceptionally low performance would not be tolerated. A core goal was to imbue a sense of professionalism into teaching in the district, making it a more attractive labor market for teachers who might have otherwise worked in local charters or neighboring districts, and making high-performing teachers, for whom IMPACT would recognize their value, more likely to stay in their schools (DCPS, 2012). Teachers would likely also be more invested in their schools if they received actionable feedback on their performance, and if their responsiveness to that feedback was incentivized by the system's sanctions and rewards (Toch, 2018).

While IMPACT includes numerous components, the core pillars of the policy include the use of multiple measures of teaching to differentiate performance, high stakes, and feedback for improvement. We briefly describe each below, beginning with the structure of IMPACT as it was first implemented in the 2009–10 school year (referred to internally as IMPACT 1.0) and 2010–11 through 2011–12 (IMPACT 2.0), followed by major revisions in 2012–13 (IMPACT 3.0).

6.3 Key Features

6.3.1 Using Multiple Measures to Differentiate Performance

Before IMPACT, the evaluation system used by DCPS, like most U.S. school districts (Weisberg et al., 2009), consisted solely of classroom observations, which were superficial, sporadic, inconsistent, and undifferentiated—rarely rating teachers less than satisfactory. IMPACT strove to provide a more rigorous and nuanced picture of effective teaching, and it did so by employing multiple measures that addressed inputs (e.g., teaching practices) and outputs (e.g., student achievement), as well as other qualities that the district felt, were important. Table 6.1 lists the measures used by IMPACT, along with the weights applied to each component, over the history of the program. The predominant measures introduced in 2009–10 were a formal classroom observation score and a score reflecting teachers' influence on their students' learning. These measures, and their respective weights, varied based on an individual's role in DCPS; in IMPACT's first year, there were 20 separate position-defined groups, each of which was subject to a distinct set of evaluation measures and procedures that were tailored to one's roles and responsibilities. This included everyone from general education teachers (groups 1 and 2), special education teachers (group 3), and teachers of English language learners (groups 4 and 5) to office staff (group 18), custodians (group 19), and other school-based personnel (group 20). We focus our attention in this chapter on general education teachers, though the core structure of IMPACT is common across many of the instructional roles.[1]

[1] These IMPACT groups have evolved over time, as DCPS has refined IMPACT and tailored the evaluation system to specific position needs. There are now more than thirty defined IMPACT groups. For a current listing of IMPACT groups, as well as detailed guidebooks describing the particulars of each group's evaluation program, visit https://dcps.dc.gov/publication/2019-2020-impact-guidebooks.

Table 6.1 IMPACT score components 2009–10 through 2019–20, by instructional role

Impact components	2009–10 (IMPACT 1.0) and 2010–11 to 2011–12 (IMPACT 2.0)		2012–13 to 2013–14 (IMPACT 3.0)		2014–15 to 2015–16 (IMPACT 3.0)	2016–17 to 2019–20 (IMPACT 3.0)	
	Teachers in tested grades and subjects	Other instructional groups	Teachers in tested grades and subjects	Other instructional groups	All teachers	Teachers in tested grades and subjects	Other instructional groups
Individual value added (IVA)	50%	N/A	35%	N/A	N/A	35%	N/A
Classroom observation	35%	75%	40%	75%	75%	30%	65%
Teacher-assessed student achievement data (TAS)	0%	10%	15%	15%	15%	15%	15%
Commitment to the school community (CSC)	10%	10%	10%	10%	10%	10%	10%
School value-added	5%	5%	N/A	N/A	N/A	N/A	N/A

(continued)

Table 6.1 (continued)

Impact components	2009–10 (IMPACT 1.0) and 2010–11 to 2011–12 (IMPACT 2.0)		2012–13 to 2013–14 (IMPACT 3.0)		2014–15 to 2015–16 (IMPACT 3.0)	2016–17 to 2019–20 (IMPACT 3.0)	
	Teachers in tested grades and subjects	Other instructional groups	Teachers in tested grades and subjects	Other instructional groups	All teachers	Teachers in tested grades and subjects	Other instructional groups
Student surveys of practice	N/A	N/A	N/A	N/A	N/A	10%	10%

Notes Group 1 consists only of those reading and mathematics teachers in grades four through eight, where it is possible to define value added with the available assessment data. IMPACT scores can also be adjusted downwards for “Core Professionalism” (CP) violations reported by principals. In the 2014–15 academic year, DCPS transitioned to a new examination for assessing student achievement; during the first two years of that transition, teachers in tested grades and subjects did not receive IVA scores, and instead had the same score components and weights as teachers in other instructional roles. The Commitment to the School Community measure (CSC) is a rubric-based assessment, scored by the school principal, of the teacher’s contributions to the professional life of the school. The Teacher-Assessed Student Achievement Data (TAS) component is a measure of student performance on a teacher-selected assessment, where performance is evaluated relative to targets set at the start of the school year; the school leader must approve both the selected measure and the teacher-developed goals. Teachers with students in grade 3 or younger are not assessed using student surveys of practice; these component weights are instead applied to the corresponding classroom observation rubric score.

DCPS's implementation of its classroom observation rubric, the Teaching and Learning Framework, was distinct from other districts in several regards. First, while many U.S. observation systems evaluate their more-experienced teachers with less frequency than novices (Steinberg & Donaldson, 2016), in the first years of IMPACT every DCPS teacher was observed five times throughout each year. Second, these observations were scored by a combination of internal raters (i.e., by school leaders, as is commonly the case), as well as by observers external to the school, known in DCPS as Master Educators, who were experienced in the same content areas and grade levels as their observers. In addition, each observation was required to be followed within two weeks by a post-observation debriefing, in which teachers would receive actionable feedback on their performance.

Teachers in tested grades and subjects (group 1) were also—and continue to be—evaluated based on their contributions to student learning, as measured by individual value-added scores. These scores essentially compare how well a teacher's students improve on the district's annual standardized assessment relative to similar students (based on prior achievement and demographic factors) with other teachers (Isenberg & Walsh, 2014). Teachers in non-tested grades and subjects are also assessed in part on their students' performance; their performance is evaluated relative to targets set at the start of the school year, with approval from the school principal on both the selected measure and the teacher-developed goals.

Additionally, DCPS included some less-common measures that reflected the district's expectations for its teaching professionals. In the first years of IMPACT, for example, DCPS included a school-level value-added measure. Teachers were also evaluated upon a rubric-based measure of their commitment to their school and community. Finally, teachers could be docked between 20 and 40 points on the IMPACT scale if they failed to demonstrate "core professionalism," such as through poor attendance, repeated late arrivals, or inadequately respectful interactions with colleagues, students, or the community.

Teachers receive a total IMPACT score based on the weighted aggregation of these measures (defined in Table 6.1). Teachers are categorized according to their scores, with the lowest-performing teachers designated "Ineffective," followed by "Minimally Effective." Teachers scoring above these thresholds were considered "Effective," with the highest-scoring teachers designated "Highly Effective." The distribution of teachers across these ratings is illustrated in Fig. 6.2.

6.3.2 High Stakes and High Rewards

IMPACT is distinguished from other teacher evaluation systems by the high stakes tied to teachers' performance. These stakes take two forms: risk of dismissal for low-performing teachers and substantial financial rewards for high-performing teachers.

The severity of performance sanctions is determined by teachers' IMPACT ratings. The lowest-performing teachers, those determined to be Ineffective based on their IMPACT score, are subject to separation prior to the next school year. Teachers who

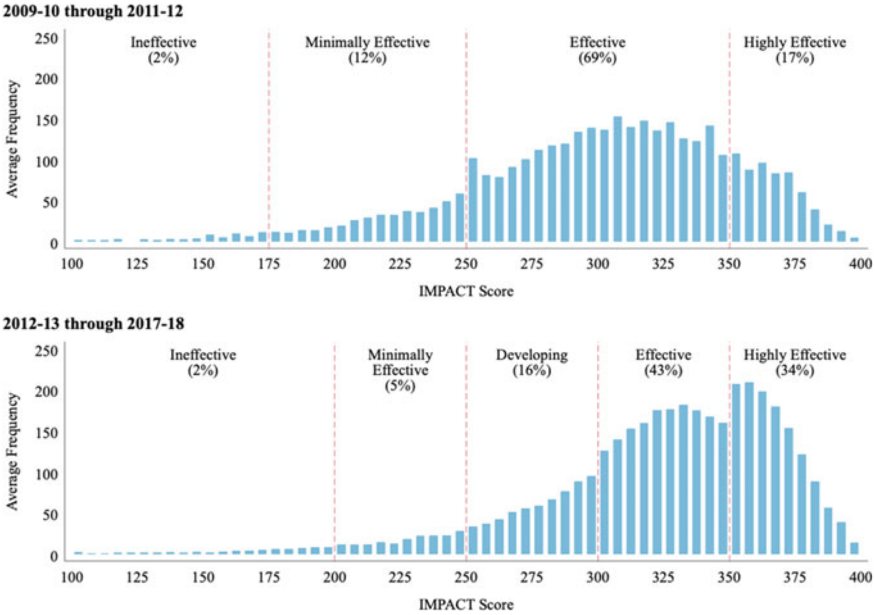


Fig. 6.2 Distribution of IMPACT performance ratings, 2009–10 through 2017–18

score above this threshold but low enough to be deemed Minimally Effective are given the opportunity to improve, but if they do not score in a higher performance band the following year, they are also subject to dismissal. Until low-performing teachers manage to score Effective or higher, they are also unable to advance on the salary scale.

Conversely, repeatedly high-performing teachers—those scoring at the Highly Effective level—can increase their base pay by skipping multiple movements across steps and lanes on the district's compensation schedule. For example, a new teacher who was repeatedly high performing over her first five years in DCPS could earn a base salary more than $60,000 above what she would have otherwise been paid (Toch, 2018). In addition, Highly Effective teachers can earn sizeable performance bonuses amounting to as much as $25,000 each year depending on their specific teaching assignments.

6.3.3 *Feedback and Improvement*

A critical component of the DCPS evaluation system is the opportunity for all but the lowest-performing teachers to improve. While each of the evaluation measures that DCPS uses provides information about how a teacher is performing, the classroom observation process is designed not only to give teachers information about how they are doing, but also as a tool for teachers to learn how they might perform better. Feedback is an explicit part of the evaluation cycle. The use of Master Educators as evaluators also facilitated this goal, as these observers have instructional expertise in the grade level and content area of the lessons that they observe, allowing for targeted feedback beyond what school administrators are able to provide. DCPS hosts a video library of teachers demonstrating exemplary practices aligned to the district's instructional standards so that teachers can view examples of effective teaching.

Critically, within fifteen days of each observation, evaluators are required to conduct a post-evaluation conference and debriefing with the teacher, in which they highlight teachers' strengths and weaknesses. They must also provide a written summary of feedback on the teacher's performance, which teachers can access through DCPS's online IMPACT portal.

6.3.4 IMPACT as an Evolving Policy

A final, but no less distinguishing, core feature of IMPACT is that DCPS has, since its inception, viewed as necessary revisions and additions to the evaluation system over time in order to maintain buy-in and effectiveness. These adaptations would be made in response to feedback from teachers and evaluators (Toch, 2018), as well as based on evaluation-induced changes to the composition of the teaching force or other contextual changes. They take the form both of explicit changes to the evaluation system, and of new policies and programs that complement or rely on IMPACT's features. An example of the former included major revisions to the observation rubric, after IMPACT's first year, to make it easier to score and more straightforward for teachers to interpret their levels of performance. The years that included and immediately followed these rubric-centered revisions were referred to by DCPS staff as IMPACT 2.0.

IMPACT 3.0. A wider sweep of changes was introduced with the start of the 2012–13 academic year.[2] This new version of IMPACT, known as IMPACT 3.0, raised expectations for effective teaching while reducing IMPACT's emphasis on value-added scores. These changes contributed to a significant reweighting of teachers' IMPACT score components (see Table 6.1) to allay teachers' concerns about the fairness and stress of being evaluated on their value-added to student achievement (Toch, 2018). Specifically, DCPS eliminated the school-level value-added measure, which teachers felt was not within their control. DCPS also reduced the weight of individual value-added scores, which had been the predominant score component for group 1 teachers and were a source of great anxiety, from 50 to 35% of their total score. The weights that had previously been applied to these measures were

[2] For a more detailed discussion of the 2012–13 changes, see Dee et al. (2019) and, for an overview of these changes in addition to other adaptations that have been made through the 2017–18 school years, Toch (2018).

reallocated in part to the district's classroom observation rubric—the Teaching and Learning Framework—which went from 35 to 40% of group 1 teachers' total score, and to teacher-assessed measures of student learning which now comprised 15% of teachers' total score.

While these changes were made in part to improve teachers' morale and support for the system, DCPS also used the performance distribution to inform some of their IMPACT 3.0 revisions (Toch, 2018). In the earliest years of IMPACT, teachers were assigned to one of four performance categories—Ineffective, Minimally Effective, Effective, and Highly Effective—yet the Effective category was so broad that it encompassed 69% of all teachers (see the top panel of Fig. 6.2). This, in combination with what was viewed as inadequate pace in the improvement of student achievement, served as a sign to DCPS that the evaluation system was not sufficiently distinguishing teacher effectiveness levels or setting appropriately high expectations for teaching. This led DCPS to redefine its score categories, making two major changes. First, the cut score for attaining a Minimally Effective rating was raised from 175 to 200. Second, what had constituted the Effective score band (250–349) now comprised two distinct ratings; teachers scoring in the lower half of the Effective range (250–299) were now considered "Developing" (see the bottom panel of Fig. 6.2). Teachers scoring at the Developing level are, like Minimally Effective teachers, unable to advance on the pay scale, but are given an opportunity to demonstrate improvements. If, however, they fail to score at least at the Effective level within the next two years—or if their score falls below Developing—they are dismissed from their jobs in DCPS.

Alongside the IMPACT 3.0 reforms, DCPS reduced the size of financial rewards for Highly Effective teachers in low-poverty schools. This change effectively shifts financial incentives toward working in—and performing well at—the schools that serve the lowest-achieving and highest-poverty students. Highly Effective teachers in low-poverty schools saw their bonus potential decrease by 75% and were no longer eligible for performance-based base pay increases. At the same time, DCPS introduced an IMPACT-aligned career ladder policy, the Leadership Initiative for Teachers (LIFT), which allowed teachers to more rapidly advance on the salary scale if they continuously scored at the Highly Effective level, in addition to providing additional recognition and leadership opportunities. As teachers advance on the performance-based career ladder, they are also subject to fewer formal classroom observations.

Other Changes to IMPACT. More recently, DCPS has implemented policies to strengthen teachers' professional development, establishing an observation, coaching, and feedback program, Learning Together to Advance Our Practice (LEAP; Cohen et al., forthcoming; Toch, 2018). This new professional development program was catalyzed by the district's concerns about teachers' readiness to teach to new, more rigorous student learning standards as DCPS transitioned to the Common Core State Standards. At the same time, policymakers were impatient with the rate of teachers' development (Toch, 2018). Under LEAP, which first rolled out in the 2016–17 school year, teachers participate in weekly group planning, professional development, and review seminars, alongside low-stakes classroom observation and feedback sessions conducted by high-performing grade- and content-aligned educators from within their schools.

A number of other changes have since been made to IMPACT, as well. With LEAP's introduction, DCPS eliminated the Master Educator evaluation program from IMPACT, given their overlapping use of grade- and content-expert evaluators, and the high cost of maintaining Master Educator-led classroom observations (Toch, 2018). DCPS also transitioned to a new classroom observation rubric, Essential Practices, which was meant to be better aligned with the Common Core State Standards and with student-centered instruction. And finally, this recent suite of changes introduced a new measure to IMPACT: Student Surveys of Practice. These student-completed surveys assess the instructional culture of the teachers' classrooms and are meant to provide teachers with actionable feedback (DCPS, 2016).

6.4 Evidence of Impact

In IMPACT's theory of change (Fig. 6.1), there are multiple dimensions across which the evaluation system in DCPS might work to improve teaching quality in the district. Some aspects of the evaluation system might promote compositional change, such as through attracting higher-quality teachers to the district via high salary potential or by encouraging the attrition of low-performing teachers by performance evaluation. Conversely, if teachers view IMPACT as onerous, it might inhibit the district's ability to recruit and hire from a rich applicant pool. Other aspects might serve to shift the distribution of teaching quality upward for the teachers who remain in DCPS, either by incentivizing the level and quality of teaching to align with the expectations defined by IMPACT, or through providing teachers with regular feedback on their performance and how they might improve. In the years since IMPACT's introduction, a number of studies have examined these mechanisms. While many contemporaneous evaluation reforms in the United States have had limited success at differentiating teaching quality or leading to meaningful effects on teaching and learning (Kraft & Gilmour, 2017; Stecher et al., 2018; Walsh et al., 2017), analyses of IMPACT indicate it has been more successful. Because IMPACT has endured over a sustained period and has incorporated meaningful changes, we provide evidence from both its early years (IMPACT 2.0) and its more recent form (IMPACT 3.0). Across both phases of IMPACT's evolution, these studies provide evidence that IMPACT has improved teaching quality and student achievement. We describe this evidence below.

6.5 Impact 2.0

Validity and Reliability. Before discussing IMPACT's effects on teaching and learning in DCPS, two first-order questions are whether IMPACT scores are reliable and valid as a measure of teaching quality. Reliability is a particular concern for a high-stakes system such as IMPACT, where misidentification due to imprecision could lead to effective teachers losing their jobs or ineffective teachers remaining in the classroom—both of which could also have meaningful consequences for students. To date, there has been no research that explicitly examines the reliability and validity of IMPACT as a whole. Multiple-measure evaluation systems, however, tend to have higher reliability than those that rely on fewer measures (Kane & Staiger, 2012). Specific to DCPS, a study from the University of Virginia (Meyer, 2016) provided estimates of the classroom observation component's reliability and validity during the first five years of IMPACT. Reliability of the observation measure was similar to or higher than other measures in the classroom observation literature (generalizability coefficients from a G-study range from 0.66 to 0.81, depending on the year and rater type (Meyer, 2016; Kane & Staiger, 2012), an important finding given that observation scores comprise a plurality of any teacher's IMPACT score.

In terms of validity, there has been a modest correlation between teachers' observation scores and their value-added scores, comparable in magnitude to what has been observed in other settings (approximately 0.30; Ho & Kane, 2013; Gill et al., 2016; Meyer, 2016; Whitehurst et al., 2014), suggesting that the two measures capture somewhat overlapping constructs. Additional evidence of construct validity comes from a study of differential teacher turnover in the district (Adnot et al., 2017). When low-performing teachers—as rated by IMPACT—exited DCPS, student achievement and teaching quality in the same grade and subject area both increased; conversely, when high-performing teachers exited, student achievement and average teaching quality each declined.

Differentiating Effective Teaching. Rigorous use of these multiple measures was expected to lead to a performance distribution that meaningfully differentiated levels of teaching quality. Indeed, IMPACT assigned a sizeable number of teachers to each of its rating categories, in contrast to many other U.S. states and school districts where, even after reforming their evaluation systems, few teachers are ever designated less than Effective (Kraft & Gilmour, 2017). More than one in seven teachers (14%) across the first three years of IMPACT was rated less than Effective (Ineffective or Minimally Effective), and a similar share (17%) was rated more than Effective (i.e., Highly Effective). DCPS, however, viewed the performance distribution as evidence that further differentiation was still needed. Through the 2011–12 academic year, 86% of teachers were rated Effective or Highly Effective, with more than two thirds of all teachers labeled Effective (Fig. 6.2, top panel). In response, and as part of the ensuing IMPACT 3.0 revisions, DCPS added another rating category, Developing, which comprised the lower half of what had initially been the Effective score band.

Recruitment. There is limited evidence on the effects of IMPACT on teacher recruitment in DCPS, but in spite of concerns more broadly that accountability reforms might deter potential teaching candidates (e.g., Kraft et al., 2020), a small handful of studies suggest that DCPS's evaluation system did not prevent the district from recruiting a large and rich applicant pool. First, Jacob et al. (2018) used applicant and hiring data from DCPS to explore the association between performance on screening measures, probability of hire, and performance on IMPACT between 2011 and 2013. They found that the screening measures, which were aligned in no small part to IMPACT and included teaching demonstrations scored on the Teaching and Learning Framework, were highly predictive of performance in the classroom, but that DCPS received far more high-performing applicants than it actually hired. Second, a paper exploring teacher turnover in the first few years of IMPACT (Adnot et al., 2017) showed that on average DCPS was able to replace its departing teachers with new hires who were at least as effective as measured both by IMPACT and by their contributions to improved student achievement.

Retention. Strategic retention represents another mechanism by which DCPS sought to improve the quality of teaching. There are two pathways by which IMPACT was designed to influence teacher retention, each based on the performance level of the teachers in question. On the lower end of the distribution, Ineffective teachers have a direct retention consequence, as they are dismissed from the district. Minimally Effective and Developing teachers can also be dismissed if they fail to adequately improve over time. These involuntary dismissals comprised approximately 3% of teachers and 16% of exits each year (author analysis). Retention effects can also occur indirectly, such as for teachers who receive low ratings that connote a threat of dismissal if they do not improve the following year and therefore preempt that possibility with a voluntary exit, or simply due to teachers making their retention decisions in response to feedback about their quality of work in the profession (Weisberg et al., 2009). Evidence from DCPS suggests that such mechanisms influenced low-performing teachers' choice to exit (Dee & Wyckoff, 2015), and that voluntary attrition comprised the vast majority of low-performing teachers' exits (Adnot et al., 2017).

Dee and Wyckoff (2015) found that earning a Minimally Effective rating had a causal effect on teachers' decisions to exit. Teachers scoring just below the Effective performance threshold were about 50% more likely to voluntarily exit DCPS than teachers just reaching that performance band (Dee & Wyckoff, 2015). The approach that Dee and Wyckoff (2015) used to evaluate these retention effects, a regression discontinuity design, makes this evidence especially compelling. This regression technique takes advantage of the similarity between a teacher who has scored just below Effective (e.g., a Minimally Effective teacher with a score of 249) and a teacher who has barely met that threshold (e.g., an Effective teacher with a score of 250); they are separated by scores that are statistically indistinguishable due to measurement error. The only difference between teachers who have just missed or just made that cut-off should be the score—and therefore the rating—itself, which allows for a clean comparison between the effect of scoring Minimally Effective (the treatment) and

Effective (the control) on teachers' retention decisions. Any difference in outcomes is therefore attributable to the rating assignment, rather than differences across teachers.

While Dee and Wyckoff found that IMPACT induced attrition among low-performing teachers, high-performing teachers may be *less* likely to exit DCPS, in response to the financial rewards for attaining a Highly Effective rating, as well as through formal recognition of their teaching quality. While the causal evidence of the financial rewards' overall retention effect is suggestive but not statistically significant (Dee & Wyckoff, 2015), high-performing (Effective and Highly Effective) teachers exited the district at far lower rates than their lower-performing (Ineffective and Minimally Effective) peers—13 v. 46% (Adnot et al., 2017).

Typically, schools and districts seek to retain as many teachers as they feasibly can, as teacher turnover can have disruptive effects on other teachers, can incur substantial time and financial costs for recruitment and hiring a replacement, and can lead to smaller achievement gains for students who are in classrooms that have been subject to turnover (Carver-Thomas & Darling-Hammond, 2019; Ingersoll, 2001; Ronfeldt et al., 2013). The differential effects of IMPACT on high- versus low-performing teachers' decisions to exit, however, belie this common wisdom. DCPS has high turnover overall, but the average turnover in DCPS between 2009 and 10 and 2011–12 led to net gains to both teaching quality and student achievement, given the diverging probabilities of exit for teachers on opposite ends of the performance spectrum, and DCPS's ability to recruit sufficiently effective replacements (Adnot et al., 2017). Student achievement improvements were substantial. When a teacher identified by IMPACT as low-performing (Ineffective or Minimally Effective) exited, teaching quality improved by 1.3 standard deviations (math) and 0.9 standard deviations (reading), and student achievement by 21% of standard deviation (math) and 14% of a standard deviation (reading).

Development. In addition to recruitment and retention, a third mechanism underlying IMPACT's design was to improve teaching quality by developing the teaching skills of existing and retained teachers in DCPS. There are multiple avenues through which IMPACT might influence teachers' skill development, but two core pathways include improving teaching quality through: (1) incenting added (or reallocated) effort and (2) feedback, both in terms of the basic receipt of information about how one is performing, as well as guidance on how to improve that performance. Several research papers on IMPACT indicate that each of these mechanisms is at play in DCPS. Phipps and Wiseman (2019), for example, used variation in the timing of unannounced classroom observations to demonstrate that the expectation of observation under IMPACT 1.0 and 2.0 led to improved performance on IMPACT, as did the feedback they received following each observation.

IMPACT's incentive effects are particularly important for encouraging teachers' development (Adnot, 2016; Dee & Wyckoff, 2015), as well as for focusing improvements to align with IMPACT's definitions of good teaching. Using the same regression discontinuity technique described previously, Dee and Wyckoff (2015) found that Minimally Effective teachers just below the Effective threshold, and therefore at risk of dismissal, in the 2010–11 academic year who nevertheless returned the next year, improved their performance by at least a quarter of a standard deviation (approximately 13 IMPACT points) relative to teachers just above the Effective threshold. The receipt of a Minimally Effective rating caused teachers to improve their score more than they would have if they had received an Effective rating. Dee and Wyckoff (2015) also found causal performance effects for Highly Effective teachers who were eligible to receive substantial bonuses upon receipt of a second Highly Effective rating. In spite of already high scores, these teachers improved by a comparable amount to their incentivized lower-performing (Minimally Effective) colleagues.

On which aspects of their teaching do incentivized teachers improve? Dee and Wyckoff (2015) provide evidence that these improvements occur across multiple IMPACT measures, including on teachers' value-added scores. Given that the classroom observation score comprises a plurality of any given DCPS teacher's overall IMPACT score, and it provides formative information—both from rubric definitions and the feedback sessions that follow each observation—about one's teaching, it may be the most salient area for teachers to focus their improvement efforts. Adnot (2016) specifically examined the improvements that incentivized teachers make on the Teaching and Learning Framework, DCPS's observation rubric at the time. Also using a regression discontinuity design, she found that Minimally Effective teachers, who must improve in order to retain their jobs, make substantially larger gains on the rubric than Effective teachers, with those gains concentrated among the most-prescriptive and the least-difficult teaching domains. Adnot's research suggests that, when incentivized, educators will alter their teaching to align with the practices for which IMPACT defines the most concrete strategies.

6.6 Impact 3.0

While early evidence demonstrated meaningful effects on teaching quality in DCPS, there are a number of reasons why we might expect the evidence on IMPACT's effects to have evolved over time. Not least among these is the degree to which IMPACT itself has evolved. IMPACT 3.0 represented significant changes to the following: (a) the set of measures, and their respective weights, upon which teachers were being evaluated; (b) the expectations for Effective performance; and (c) the consequences associated with a given teacher's performance level. At the same time, the context within which IMPACT was operating had also evolved over the course of the program's life. DCPS has undergone multiple transitions in leadership, support for high-stakes teacher evaluation has waned nationally, and the composition of the teaching force has changed at least in part due to IMPACT itself, as described above (Dee et al., 2019). It is also possible that internal pressures and changing priorities might have begun to attenuate or reverse its positive effects. More recent evidence, however, suggests that IMPACT's measurement properties have largely held over time and that its effects on teacher recruitment, retention, and development have remained remarkably resilient.

Validity and Reliability. While there is no evidence across the full span of recent years about IMPACT's statistical reliability, the University of Virginia reliability and validity study (Meyer, 2016) covered the time from the start of IMPACT (2009–10) through the first two years of IMPACT 3.0 (ending in 2013–14). The yearly generalizability coefficients for the Teaching and Learning Framework observation scores were nearly constant across this period, indicating that the reliability of IMPACT's dominant measure did not decline over time.[3]

In terms of validity, confirmatory evidence of continuing construct validity comes from a recent study of differential teacher turnover in the district (James & Wyckoff, 2020) which closely follows the earlier methods and findings of Adnot et al. (2017). Student achievement and teaching quality each increase following the exit of a low-rated teacher, while exits of highly rated teachers lead to declines in student achievement and teaching quality. These effects are somewhat smaller than the earlier analysis, which may be anticipated since the stock of lower-performing teachers has been reduced, either through attrition or improvement.

Importantly, recent surveys of DCPS teachers indicate general agreement with the face validity of IMPACT. A majority of teachers report having somewhat to strong agreement with the performance criteria used to evaluate their teaching and the accuracy of their performance ratings as reflections of their effectiveness (69 and 65%, respectively; James & Wyckoff, 2020).

Differentiating Effective Teaching. As part of the suite of changes to IMPACT implemented in 2012–13, DCPS bisected its previously defined Effective score range such that teachers scoring in the lower half of that score band were now considered Developing rather than Effective. This change was made in response to a perception that the large share of teachers scoring at the Effective level (69%) was evidence of insufficient performance differentiation. This change redistributed the share of teachers who were considered high- versus low-performing (see the second panel of Fig. 6.2). Across years since the score-band revision, close to a quarter (23%) of teachers are rated less than Effective (i.e., Ineffective, Minimally Effective, or Developing), although the share of high-performing teachers has trended upward over time (Dee et al., 2019; James & Wyckoff, 2020).

Recruitment. The evidence on IMPACT and recruitment remains thin. A recent paper exploring teacher turnover under IMPACT 3.0 (James & Wyckoff, 2020), however, produces results closely aligned to Adnot et al.'s (2017) findings from IMPACT 2.0; on average, DCPS is able to replace its departing teachers with new hires who are at least as effective as measured both by IMPACT scores and by their contributions to student achievement. This is in spite of increases to the average effectiveness of exiting teachers as the district's performance distribution shifted upward over time (James & Wyckoff, 2020).

The incentive structure built into IMPACT 3.0, where Highly Effective teachers can earn larger bonuses and advance more rapidly on the career ladder if they work in

[3] Across evaluators and instructional groups, reliability values from the G-study were 0.72 in 2009–10 and 0.74 in 2013–14.

high-poverty schools (i.e., schools that disproportionately serve students from low-income families), may contribute to successful recruitment within the district.[4] Katz and Wiseman (2020) find that, following the reforms that assigned greater financial rewards to high-poverty schools, Highly Effective teachers in DCPS were more likely to transfer from low- to high-poverty schools. Katz and Wiseman also find suggestive evidence that high-poverty schools were able to improve their hiring yield (i.e., to increase the share of new teachers who were Effective or higher) as a result of these financial incentives.

Retention. Early evidence from DCPS demonstrated that the incentive mechanisms that were built into IMPACT induced low-performing teachers to exit (Dee & Wyckoff, 2015) and that most of this attrition was voluntary (Adnot et al., 2017). In more recent research on retention effects under IMPACT 3.0, Dee et al. (2019) found nearly identical retention effects for Minimally Effective teachers to what Dee and Wyckoff (2015) had observed under IMPACT 2.0. Earning a Minimally Effective rating caused higher rates of attrition than earning the next-higher rating; teachers scoring just below the Developing performance threshold (i.e., Minimally Effective) were about 50% more likely to voluntarily exit DCPS than teachers just reaching that performance band (Dee et al., 2019). IMPACT 3.0 also included a new rating category under which teachers were subject to a similar, albeit less immediate, dismissal threat. Retention effects for teachers who had received their first Developing rating were nearly as large; these teachers experienced an increase in voluntary attrition of approximately 40%.

While we do not have more recent causal estimates of retention effects for high-performing teachers, James and Wyckoff (2020) demonstrate that high-performing teachers continue to exit the district at far lower rates than their lower-performing peers (Adnot et al., 2017; James & Wyckoff, 2020). In addition, the differentially large pay and bonus incentives for teaching in high-poverty schools lead to higher retention among Highly Effective teachers in the schools that need them the most (Katz & Wiseman, 2020).

These sustained retention effects over IMPACT's evolution do not inherently imply continued benefits to students from differential turnover (Adnot et al., 2017). IMPACT has by design incentivized (or compelled, in the case of very low-performing teachers) the exit of less-effective teachers, while also incentivizing and aiming to facilitate the development of those who remain (Adnot, 2016; Dee & Wyckoff, 2015; Dee et al., 2017). This has led to fewer teachers being rated less than Effective over time—a trend that corresponds to the average exiting teacher being of higher quality in more recent years than in earlier years. If DCPS loses more Effective or Highly Effective teachers than it is able to hire, attrition in DCPS might lead to negative consequences for overall teaching quality and for student achievement. James and Wyckoff (2020), however, find that the average turnover in DCPS continues to provide net gains to both teaching quality and student achievement, as

[4] DCPS defines high-poverty schools as schools that serve high proportions of students in low-income families; specifically, high-poverty schools are those in which 60% or more of the student body is eligible for free or reduced-price lunch.

DCPS still observes diverging probabilities of exit for teachers on opposite ends of the performance spectrum, and is still able to recruit sufficiently effective replacements (James & Wyckoff, 2020). While these net effects have diminished over time as the composition of teachers in the district has improved, they remain substantial.

Development. Adnot (2016) and Dee and Wyckoff (2015) showed that incentive effects during IMPACT 2.0 were particularly important for encouraging teachers' development, as well as for focusing improvements to align with IMPACT's definitions of good teaching. Dee et al. (2019) revisited this question using data from IMPACT 3.0. They found little indication of performance effects for Developing teachers. However, Minimally Effective teachers just below the Developing threshold, and therefore at risk of dismissal in the subsequent academic year, who nevertheless returned the next year improved their performance by roughly a quarter of a standard deviation (approximately 13 IMPACT points) relative to teachers just above the Developing threshold—comparable to the performance effects that Dee and Wyckoff (2015) had observed in IMPACT 2.0. The authors were unable to determine which of IMPACT's measures were driving teachers' improvements (i.e., the sub-measure effects were statistically insignificant), but their analyses suggested that Minimally Effective teachers made substantial gains on IMPACT's more formative measures—improving on their rubric-assessed classroom observation scores (the Teaching and Learning Framework) and on their Commitment to their School and Community, consistent with Adnot's (2016) earlier evidence that incentivized teachers improve on measures that provide descriptions of exemplary practice.

6.7 Conclusions

DCPS introduced IMPACT during a time when teacher evaluation reforms were being implemented across the United States. Even PERAA, which paved the way for IMPACT, was motivated by effective efforts in other urban districts facing similar challenges. These districts, including Boston, Chicago, and New York City, introduced reforms that relied on data to make education policy decisions and emphasized a culture of learning and improvement. These exemplar districts also turned over control of their public schools to their respective mayors. By following suit, DC hoped that it could similarly offer its public education system the change needed to improve student outcomes (NRC, 2011). Similarly, DCPS was not alone in making teacher evaluation a centerpiece of its school reforms, as a growing body of evidence pointed to the importance of teachers for an array of student outcomes (Aaronson et al., 2007; Chetty et al., 2014a, 2014b; Kane & Staiger, 2008; Rivkin et al., 2005; Rockoff, 2004; Sanders & Rivers, 1996). At the same time, many evaluation systems failed to effectively distinguish levels of teaching quality, which inhibited districts' ability to facilitate teacher development or to hold ineffective teachers accountable

(Weisberg et al., 2009). Spurred in large part by federal grant programs that incentivized rigorous teacher evaluation, districts, and states across the USA rapidly implemented teacher evaluation reforms over the course of a brief period (Steinberg & Donaldson, 2016).

DCPS is not the only place in the USA to have developed an evaluation system with evidence of benefits to teaching and learning. Chicago (Sartain & Steinberg, 2016; Steinberg & Sartain, 2015), Cincinnati (Taylor & Tyler, 2012), and Houston (Cullen et al., 2017) each have produced evidence of successful evaluation programs. At the same time, however, a number of places have seen little or no effects from their reform efforts. In spite of significant changes to evaluation systems across the United States, the typical district still fails to meaningfully differentiate levels of teaching quality (Kraft & Gilmour, 2017). Large-scale and high-profile interventions to reform teacher evaluation in other parts of the USA have demonstrated null or mixed effects on student outcomes and teaching quality (e.g., Stecher et al., 2018). The reasons for this are complex, but researchers have pointed to incomplete implementation and competing policies as likely factors (e.g., Cullen et al., 2019; Stecher et al., 2018).

What explains DCPS's success, while many other locations across the USA have struggled to implement and maintain rigorous teacher evaluation programs? We suspect that many factors contributed. One is the unusual governance structure for the District of Columbia which allowed the U.S. Congress to cede control of teacher evaluation in the district to the local school board, outside of the typical bargaining between the school board and the teachers' union (Toch, 2018).

DCPS leadership appears to privilege the urgent needs of the children in the district while addressing the needs and preferences of its teachers. This includes the revisions that have been made to IMPACT as the district adjusts to changing contexts and evidence of the program's strengths and weaknesses. DCPS has likewise been careful to make design decisions that increase or strengthen teachers' buy-in to the policy; for example, every school-based employee in DCPS is evaluated on IMPACT, rather than just teachers (or teachers and school leaders). These efforts may have bolstered teachers' perceptions of IMPACT; in recent surveys administered across the district, most teachers indicate agreement that the teacher evaluation process helps them identify their strengths and weaknesses, that IMPACT ratings are accurate reflections of teacher effectiveness, and that they agree with the criteria used to evaluate their performance (James & Wyckoff, 2020). Support of the evaluation systems is particularly high among Highly Effective teachers, for whom buy-in is most critical. Finally, DCPS is fortunate to have a rich and deep applicant pool from which it can hire new teachers—a feature that has enabled DCPS to mitigate the negative effects of teacher turnover in the district.

DCPS, however, is not immune to the pressures that have caused other evaluation reform attempts to falter. As IMPACT enters its second decade, its future is uncertain. In spite of its successes, sustaining IMPACT remains a complex endeavor. First, implementing and maintaining a robust teacher evaluation system can be expensive, requiring infrastructure, additional staffing, and training (Donaldson & Papay, 2015; Stecher et al., 2018). IMPACT's large financial rewards add additional costs. In the 2019 fiscal year, for example, DCPS spent $2.2 million on maintaining IMPACT and more than $15 million on teachers' performance-based bonuses (DCPS, 2020a). As the average performance level of teachers in DCPS has increased over time, so too have the costs of paying out financial rewards to Highly Effective teachers. IMPACT likewise has continued to face political pressure to remove its most binding features or for the program to be eliminated entirely (Stein, 2019a, 2019b). In June 2019, the City Council introduced legislation that would give bargaining power over teacher evaluation to the teachers' union while also prohibiting negative consequences from evaluations that have not been agreed to through collective bargaining. If this bill were to pass, it could very well end IMPACT, as the local union leadership is vocal in their opposition to the policy (Stein, 2019a). Representatives from the teachers' union are openly critical of IMPACT's performance sanctions and the anxiety that IMPACT's high stakes impart upon teachers in the district. Meanwhile, the current chancellor of DCPS, Lewis D. Ferebee, recently initiated an in-depth study of the first ten years of the program (DCPS, 2020b). While Ferebee does not intend for this evaluation to lead to a complete redesign of IMPACT, additional changes are likely (Stein, 2019b).

Implicit in Ferebee's decision to conduct this review of IMPACT is the recognition that, in spite of IMPACT's success, there remains considerable room for improvement (DCPS, 2020b). While DCPS has made large gains to student achievement over the past decade, achievement gaps remain stubbornly wide across students from different demographic backgrounds. For example, fewer than one in five (18.1%) Black or African-American students were defined as meeting or exceeding expectations based on their scores from the most recent district-wide standardized assessment (OSSE, 2019). Outcomes were somewhat better (33.9%) for Hispanic or Latino students, but substantially higher for white students (82.1%). Nearly identical gaps in proficiency levels are observed on other measures, including the 2019 National Assessment of Educational Progress (USDOE, 2019). While there is evidence that differential incentives for high performance in high- versus low-poverty schools encourage higher retention, increase transfers from low-poverty schools, and possibly result in more effective recruitment in the schools that are in the greatest need of high-performing teachers, there is still an inequitable distribution of Highly Effective teachers across schools. In 2018, for example, the average high-poverty school had approximately a third (33%) of its teachers earn a Highly Effective rating, while the average low-poverty school had more than half of its teaching staff score at the Highly Effective level (57%; author analysis). In addition, while DCPS retains substantially higher proportions of its highest-performing teachers than its low performers, there may be room for improved retention even for DCPS's most-effective teachers, with approximately one in ten Highly Effective teachers leaving each year (James & Wyckoff, 2020).

Whatever form IMPACT takes in the near future, a number of questions remain unanswered. Generally, are there ways to make IMPACT more supportive for teachers (e.g., less stressful) without mitigating its benefits for teacher quality and student learning (DCPS, 2020)? Questions that could be addressed include:

- To what extent are high stakes necessary in order to incentivize teachers' improvement or their differential retention? Could less onerous sanctions achieve comparable results at less political cost?
- What are the trade-offs between unannounced and scheduled classroom observations? Might teachers be more receptive to feedback, would they perform better in the absence of uncertainty about whether they are going to be evaluated each day, or would teachers be more likely to manipulate their teaching in undesirable ways if they knew in advance when an evaluation was going to occur?
- Do any features of the system result in unintended consequences? One feature that might drive unintended consequences is the pay-scale advancement process; for example, teachers who are rated Developing are unable to advance on the salary scale even if they have improved from a lower (i.e., Minimally Effective) rating. Do these step holds drive teachers' behavior in a way that inhibits their performance? It is also possible that the strong stakes induce teachers to manipulate IMPACT measures, for example, by creating a "lesson in a box" ready for deployment when an unannounced classroom observation occurs. To the extent this misrepresents a teacher's underlying true teaching performance, weaknesses may go unaddressed. Other unintended consequences might include teachers distorting their effort across different teaching tasks and measures, including in ways that have negative effects on students' non-academic outcomes. Experienced teachers might also move to teaching assignments in classrooms, subjects, or grades where they expect it to be easier to score well on IMPACT.
- We also know little about the effectiveness of teacher recruitment in DCPS and the extent to which IMPACT might facilitate or inhibit the district's staffing. Does the risk of dismissal deter candidates from applying or does the rigor of evaluation or high salary potential attract applicants looking for more professionalized teaching contexts? Are these effects constant across applicant skill levels or teaching contexts?
- IMPACT has undergone a number of changes—not simply in terms of the measures that comprise IMPACT, but also the ways in which it is implemented, and the teaching population to which it is applied. What are the measurement properties of the current iteration? How might reliability and validity be improved?
- Recent research on evaluation systems has shown that bias on the part of the evaluator or in the rubric itself may influence the scores that teachers receive (e.g., Chi, 2020; Steinberg & Sartain, 2020). To what extent does implicit bias affect classroom observations, and what could DCPS do to address this?
- What are the trade-offs associated with DCPS's transition away from independent evaluators? For example, while the Master Educator program was costly (Toch, 2018), did its elimination reduced the accuracy of scoring—either in terms measurement properties (e.g., reliability or alignment to other, more objective measures) or teacher's perceptions of the validity of their scores?
- While IMPACT directly affects some teachers' retention decisions, and Highly Effective teachers are less likely to exit DCPS than their less-effective peers, turnover of any highly effective teacher comes at a cost to students. Are there additional ways IMPACT could be used to support the most-effective teachers'

retention in DCPS, such as through additional opportunities for growth, resources to lighten workloads, or efforts to facilitate more supportive school leadership?

- Finally, a core question about DCPS's continuing ability to sustain IMPACT is its cost. Are there less expensive ways of implementing IMPACT that would preserve its most-effective features? Could smaller financial incentives, for example, generate comparably large effects?

Answers to each of these questions are crucial for IMPACT moving forward. To date, the accumulating evidence indicates that teacher evaluation as implemented in DCPS has led to meaningful improvements to the teaching force that extend to the district's students. These gains are, however, tenuous; they rely on careful and thoughtful implementation that is continuously responsive to the political context in which it operates, information about its successes and failures, and—most importantly—the needs of its teachers and students.

References

Aaronson, D., Barrow, L., & Sander, W. (2007). Teachers and student achievement in the Chicago public schools. *Journal of Labor Economics, 25*(1), 95–135.

Adnot, M. K. (2016). *Effects of incentives and feedback on instructional practice: Evidence from the District of Columbia Public Schools' IMPACT teacher evaluation system* (Doctoral Dissertation, University of Virginia).

Adnot, M., Dee, T., Katz, V., & Wyckoff, J. (2017). Teacher turnover, teacher quality, and student achievement in DCPS. *Educational Evaluation and Policy Analysis, 39*(1), 54–76.

Boyd, D., Lankford, H., Loeb, S., Ronfeldt, M., & Wyckoff, J. (2010). The role of teacher quality in retention and hiring: Using applications to transfer to uncover preferences of teachers and schools. *Journal of Policy Analysis and Management, 30*(1), 88–110.

Boyd, D., Lankford, H., Loeb, S., & Wyckoff, J. (2013). Analyzing the determinants of the matching of public school teachers to jobs: Disentangling the preferences of teachers and employers. *Journal of Labor Economics, 31*(1), 83–117.

Carver-Thomas, D., & Darling-Hammond, L. (2019). The trouble with teacher turnover: How teacher attrition affects students and schools. *Education Policy Analysis Archives, 27*(36), 1–27.

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014a). Measuring the impacts of teachers 1: Evaluating bias in teacher value-added estimates. *American Economic Review, 104*(9), 2593–2632.

Chetty, R., Friedman, J. H., & Rockoff, J. E. (2014b). Measuring the impacts of teachers II: Teacher value-added and student outcomes in adulthood. *American Economic Review, 104*(9), 2633–2679.

Chi, O. L. 2020. *A classroom observer like me: The effect of demographic congruence between teachers and raters on observation scores.*

Cohen, J. C., Wyckoff, J., Katz, V., Boguslav, A., Sadowski, K., & Wiseman, E. A. (Forthcoming). Implementing targeted professional development at scale in the District of Columbia Public Schools. *American Educational Research Journal.*

Dee, T. S., James, J., & Wyckoff, J. (2019). Is effective teacher evaluation sustainable? Evidence from DCPS. *Education Finance and Policy.*

Dee, T. S., & Wyckoff, J. (2015). Incentives, selection, and teacher performance: Evidence from IMPACT. *Journal of Policy Analysis and Management, 34*(2), 267–297.

District of Columbia Public Schools [DCPS]. (2012). *An overview of IMPACT.* Washington, DC: Author. Retrieved from: https://web.archive.org/web/20120118221026/; http://dcps.dc.gov/DCPS/In+the+Classroom/Ensuring+Teacher+Success/IMPACT+%28Performance+Assessment%29/An+Overview+of+IMPACT

District of Columbia Public Schools [DCPS]. (2016). *2016–2017 IMPACT guidebooks.* Washington, DC: Author. Retrieved from: https://dcps.dc.gov/publication/2016-2017-impact-guidebooks

District of Columbia Public Schools [DCPS]. (2020a). *Responses to FY2019 performance oversight questions.* Washington, DC: Author. Retrieved from https://dccouncil.us/wp-content/uploads/2020/02/dcps_Part1.pdf

District of Columbia Public Schools [DCPS]. (2020b, January 24). *American University partners with DCPS to research the IMPACT teacher evaluation system.* Retrieved from https://dcps.dc.gov/release/american-university-partners-dcps-research-impact-teacher-evaluation-system

Donaldson, M. L., & Papay, J. P. (2015). Teacher evaluation for accountability and development. In H. F. Ladd & M. E. Goertz (Eds.), *Handbook of research in education finance and policy* (pp. 174–193). Routledge.

Gill, B., Shoji, M., Coen, T., & Place, K. (2016). *The content, predictive power, and potential bias in five widely used teacher observation instruments* (REL 2017–191). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Mid-Atlantic.

Guarino, C. M., Brown, A. B., & Wyse, A. E. (2011). Can districts keep good teachers in the schools that need them most? *Economics of Education Review, 30*(5), 962–979.

Ho, A. D., & Kane, T. J. (2013). *The reliability of classroom observations by school personnel.* Seattle, WA: Measures of Effective Teaching Project, Bill & Melinda Gates Foundation.

Ingersoll, R. M. (2001). Teacher turnover and teacher shortages: An organizational analysis. *American Educational Research Journal, 38*(3), 499–534.

Isenberg, E., & Walsh, E. (2014). *Final report: Measuring teacher value added in DC, 2013–2014 school year.* Washington, DC: Mathematica Policy Research.

Jackson, C. K. (2009). Student demographics, teacher sorting, and teacher quality: Evidence from the end of school desegregation. *Journal of Labor Economics, 27*(2), 213–256.

Jackson, C. K. (2013). Match quality, worker productivity, and worker mobility: Direct evidence from teachers. *Review of Economics and Statistics, 95*(4), 1096–1116.

Jacob, B. A., Rockoff, J. E., Taylor, E. S., Lindy, B., & Rosen, R. (2018). Teacher applicant hiring and teacher performance: Evidence from DC public schools. *Journal of Public Economics, 166*(1), 81–97.

James, J., & Wyckoff, J. (2020). Teacher evaluation and teacher turnover in equilibrium: Evidence from DC public schools. *AERA Open, 6*(2), 1–20. https://doi.org/10.1177/2332858420932235

Kane, T. J., & Staiger, D. O. (2008). *Estimating teacher impacts on student achievement: An experimental evaluation* (NBER Working Paper 14607). Cambridge, MA: National Bureau of Economic Research. Retrieved from the National Bureau of Economic Research: http://www.nber.org/papers/w14607

Kane, T. J., & Staiger, D. O. (2012). *Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains.* Seattle, WA: The Bill & Melinda Gates Foundation.

Katz., V. L., & Wiseman, E. A. (2020). *Using financial incentives to attract and retain high-performing teachers in low-performing schools: Evidence from D.C. Public Schools.* Working Paper.

Kraft, M. A., & Gilmour, A. (2017). Revisiting *The Widget Effect*: Teacher evaluation reform and the distribution of teacher effectiveness. *Educational Researcher, 46*(5), 234–249.

Kraft, M., Brunner, E., Dougherty, S., & Schwegman, D. (2020). Teacher accountability reforms and the supply and quality of new teachers. *Journal of Public Economics, 188*(1), 1–24.

Meyer, J. P. (2016). *Reliability of and validity evidence for teaching learning framework scores for the district of Columbia public school system.* Unpublished manuscript, Curry School of Education, University of Virginia, Charlottesville, VA.

National Research Council [NRC]. (2011). *A plan for evaluating the district of Columbia's public schools: From impressions to evidence*. The National Academies Press.

National Research Council [NRC]. (2015). *An evaluation of the public schools of the district of Columbia: Reform in a changing landscape*. The National Academies Press.

Office of the State Superintendent of Education [OSSE]. (2019). Detailed 2018–19, 2017–18, 2016–17 PARCC and MSAA performance [Excel workbook]. Retrieved from https://osse.dc.gov/page/2018-19-parcc-results-and-resources

Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, schools, and academic achievement. *Econometrica, 73*(2), 417–458.

Rockoff, J. E. (2004). The impact of individual teachers on student achievement: Evidence from panel data. *American Economic Review, 94*(2), 247–252.

Ronfeldt, M., Loeb, S., & Wyckoff, J. (2013). How teacher turnover harms student achievement. *American Educational Research Journal, 50*(1), 4–36.

Sanders, W. L., & Rivers, J. C. (1996). *Cumulative and residual effects of teachers on future student academic achievement*. University of Tennessee Value-Added Research and Assessment Center.

Sartain, L., & Steinberg, M. P. (2016). Teachers' labor market responses to performance evaluation reform: Experimental evidence from Chicago public schools. *Journal of Human Resources, 51*(3), 615–655.

Stecher, B. M., Holtzman, D. J., Garet, M. S., Hamilton, L. S., Engberg, J., Steiner, E. D., Chamber, J., et al. (2018). *Improving teaching effectiveness: Final report: The Intensive Partnerships for effective teaching through 2015–16*. Santa Monica, CA: RAND Corporation.

Steinberg, M. P., & Sartain, L. (2020). What explains the race gap in teacher performance ratings? Evidence from Chicago public schools. *Educational Evaluation and Policy Analysis*, 1–23.

Stein, P. (2019a, June 30). With union backing, D.C. Council introduces proposed overhaul of controversial teacher evaluation system. *The Washington Post.* https://www.washingtonpost.com/local/education/with-union-backing-dc-council-introduces-proposed-overhaul-of-controversial-teacher-evaluation-system/2019a/06/29/f3722a7a-992f-11e9-8d0a-5edd7e2025b1_story.html

Stein, P. (2019b, October 21). Chancellor pledges to review D.C.'s controversial teacher evaluation system. *The Washington Post*. https://www.washingtonpost.com/local/education/chancellor-vows-to-review-the-districts-controver-sial-teacher-evaluation-system/2019b/10/20/6c00405c-f0de-11e9-8693-f487e46784aa_story.html

Steinberg, M. P., & Donaldson, M. L. (2016). The new educational accountability: Understanding the landscape of teacher evaluation in the post-NCLB era. *Education Finance and Policy, 11*(3), 340–359.

Steinberg, M. P., & Sartain, L. (2015). Does teacher evaluation improve school performance? Experimental evidence from Chicago's excellence in teaching project. *Education Finance and Policy, 10*(4), 535–572.

Taylor, E. S., & Tyler, J. H. (2012). The effect of evaluation on teacher performance. *American Economic Review, 102*(7), 3628–3651.

Toch, T. (2018). *A policymaker's playbook: Transforming public school teaching in the nation's capital.* Washington, DC: Future Ed, Georgetown University https://www.future-ed.org/wp-content/uploads/2018/06/APOLICYMAKERSPLAYBOOK.pdf

U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP). (2007). Reading and Math Assessments. Reports generated using the NAEP Data Explorer. http://nces.ed.gov/nationsreportcard/naepdata

U.S. Department of Education [USDOE], Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP). (2019). Reading and Math Assessments. Reports generated using the NAEP Data Explorer. http://nces.ed.gov/nationsreportcard/naepdata

Walsh, K., Joseph, N., Lakis, K., & Lubell, S. (2017). *Running in place: How new teacher evaluations fail to live up to promises.* Washington, DC: National Council on Teacher Quality. Retrieved from https://www.nctq.org/publications/Running-in-Place:-How-New-Teacher-Evaluations-Fail-to-Live-Up-to-Promises

Wayne, A. J., Yoon, K. S., Zhu, P., Cronen, S., & Garet, M. S. (2008). Experimenting with teacher professional development: Motives and methods. *Educational Researcher, 37*(8), 469–479.

Weisberg, D., Sexton, S., Mulhern, J., & Keeling, D. (2009). *The widget effect: Our national failure to acknowledge and act on differences in teacher effectiveness.* New York, NY: The New Teacher Project.

Whitehurst, G. J. R., Chingos, M. M., & Lindquist, K. M. (2014). *Evaluating teachers with classroom observations.* Washington, DC: Brown Center on Education Policy: Brookings Institute.

Yoon, K. S., Duncan, T., Lee, S. W. Y., Scarloss, B., & Schapley, K. (2007). *Reviewing the evidence on how teacher professional development affects student achievement* (Issues & Answers Report, REL 2007—No. 033). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Southwest.

Chapter 7
From Formulation to Impact: Lessons Learned from Teacher Evaluation Reform in Tennessee, USA

Luis A. Rodriguez

Abstract Teacher evaluation has been a considerable focus across many countries around the globe. The concern for how teachers are evaluated is largely predicated on decades of research that suggests high-quality teaching is crucial for student learning. In attempt to improve teaching quality, education policymakers have advanced various reforms to systems designed to evaluating teacher performance. This chapter considers the case of the teacher evaluation reform in Tennessee. In 2011, the Tennessee state legislature voted to implement a number of changes to the teacher evaluation process, which—similar to many other states at the time within the United States—were enacted in attempt to compete for federal education grants under the Race to the Top grant competition. This chapter offers an in-depth recounting of the formulation, passage, and implementation of Tennessee's statewide teacher evaluation reforms. In addition, the chapter outlines the major design elements of the reformed teacher evaluation system, its underlying theory of action, and its impact as documented in the extant research base.

7.1 Introduction

Located in the heartland of the politically conservative Southern United States, Tennessee has a long tradition of espousing the implementation and study of education reforms. From the class size experiment, Project Student–Teacher Achievement Ratio (STAR) in the 1980s to the Project on Incentives in Teaching (POINT) in the late-aught, the Tennessee Department of Education (TDOE) has been among the few state and district education agencies to pioneer some of the most comprehensive education reforms to ever take hold both within and outside of US borders. In the light of the state's history with education reform and experimentation, one may consider it unsurprising that in the last decade, Tennessee has aggressively pursued

L. A. Rodriguez (✉)
New York University, New York City, NY, USA
e-mail: luis.a.rodriguez@nyu.edu

J. Manzi et al. (eds.), *Teacher Evaluation Around the World*,
Teacher Education, Learning Innovation and Accountability,
https://doi.org/10.1007/978-3-031-13639-9_7

and enacted comprehensive policy reforms affecting the K-12 teacher working force, with teacher evaluation reform functioning as their centerpiece.

At the time of this writing, the TDOE comprises over sixty-five thousand public school teachers serving over one million students (Glander, 2017). Under current state policy, each teacher—with some exception[1]—is required to be evaluated annually based on multiple metrics of performance that are based in part on measured student achievement and growth. The annual teacher evaluation process is theorized to have tremendous influence on teachers, instruction, and students. First and foremost, and as codified in legislative language, the primary purpose of annual evaluation is to identify and support instruction through formative feedback while informing human capital decisions, such as individual group professional development plans, hiring, assignment and promotion, tenure and dismissal, and compensation (Tennessee State Board of Education, 2020a). In this regard, teacher evaluation reform has become an integral part of the teaching profession and promises to remain so for years to come.

As one of the most widely studied education contexts both within the United States and abroad, Tennessee provides a rich source of information on how a fully implemented teacher evaluation reform unfolds, specifically when moderate stakes for teachers are embedded in the fabric of its design. While the checkered history and effects of evaluation reform in Tennessee have been chronicled across many journalistic and academic sources, attempts to comprehensively synthesize how the system came to be and what lessons it has to offer have been few and far between. The chapter presented herein marks a renewed attempt to contextualize one of the most broad statewide education reforms within the history of the United States, one that not only has affected every single teacher within the educational jurisdiction of the state but that has also affected the teaching profession more broadly by serving as a model for teacher evaluation in other educational settings.

The chapter considers the case of the reformed teacher evaluation system in Tennessee by first providing an in-depth recounting of its formulation and passage at the state legislature, with a focus on the political context and manner in which the evaluation reform was enacted. It then continues by expounding on the major elements of the reformed evaluation system, its underlying theory of action, and the evidence base supporting its validity. Next, the chapter presents the documented ways in which stakeholders perceive and have leveraged the system to improve teacher development and staffing as well as the major implementation challenges confronting the system since its inception and the ways in which the system has been modified to address such challenges. The final sections present the estimated effects of the reformed teacher evaluation system on teacher satisfaction, performance, and retention, and other outcomes of interest. The chapter concludes by considering lessons learned from the Tennessee context that may credibly translate to the broader fields of education policymaking and education administration.

[1] Tennessee statute stipulates that evaluation requirements do not apply to teachers who are employed full time for less than 120 school days or are not employed full time—these teachers are granted a partial year exemption (PYE). Reasons for PYE include teachers with contracts less than 120 days, teachers who have been on extended leave, and teachers who transferred to a different school during the academic year (Tennessee Department of Education, 2013).

7.2 Formulation and Passage of Teacher Evaluation Reform

> I want to praise Tennessee's continuing effort to improve support and evaluation for teachers. For too long, in too many places, schools systems have hurt students by treating every teacher the same—failing to identify those who need support and those whose work deserves particular recognition. Tennessee has been a leader in developing systems that do better—and that have earned the support of a growing number of teachers.
>
> Former U.S. Secretary of Education Arne Duncan, U.S. Department of Education (2013)

Perhaps since 1992, the writing was on the wall that Tennessee would someday pioneer an innovative teacher evaluation system, for that was the year the state laid the foundation to link student academic outcomes to educational evaluation. On the heels of passing a major increase in funding for education in the state—and an increase in the state's sales tax rate to finance that boost in funding—state legislators demanded strong accountability provisions to ensure that the new funds would be invested wisely and lead to improvements in student achievement (Sanders & Horn, 1998). Among the many measures that formed the base for the state's new educational accountability system, the Tennessee Value-Added Assessment System (TVAAS), was perhaps the most innovative and controversial. Based on a statistical methodology designed to ascertain the effectiveness of school systems, schools, and teachers in producing student academic gains, TVAAS would later become one of the centerpieces of teacher evaluation reform nearly two decades later.

In July 2009, then President Barack Obama and Secretary of Education Arne Duncan announced the start of the "Race to the Top" (RTTT) program, one of the biggest and most innovative US federal investments in education to date. Funded as part of the economic stimulus package under American Recovery and Reinvestment Act of 2009, RTTT allocated $4.35 billion to support a competitive grant process intended to spur and reward education innovation and reforms at the state and local district level. RTTT involved an inducements-with-competition approach to enact education policy reform that aligned with federal policy aims. To successfully apply for RTTT funds, states submitted proposals outlining a set of prospective educational policies and standards that aligned with a set of established selection criteria, among which involved the implementation of a comprehensive educator evaluation system for teachers and school leaders (U.S. Department of Education, 2009a). Evaluation systems deliberately designed to inform human capital decisions, including but not limited to individual and group professional development plans, hiring, assignment and promotion, compensation, and tenure and dismissal further enhanced a state's prospect of receiving RTTT funds. Among the 40 states that submitted letters of intent to apply for Phase 1 of RTTT, Tennessee's vision for educational reform gained favor, as it was named one of the first two states to win a RTTT grant—the other being Delaware—worth $500 million (U.S. Department of Education, 2009b, 2010a). To receive the funds, the state legislature—with the support of then Governor Phil

Bredesen—rewrote key education provisions largely in line with the array of education reforms proposed in the state's RTTT application, which succeeding governor, Bill Haslam, later signed into law.

The passage of the RTTT-inspired education reforms was emblematic of a rare two-party political coalition and widespread statewide buy-in for comprehensive plans to reform schools in Tennessee. Not only did the RTTT application effort and subsequently legislated reforms receive support from both an outgoing liberal-leaning governor and newly elected conservative governor, but the reforms were passed with near-unanimous support from Democratic and Republican state legislators in a matter of days. After introducing the bill on January 12, 2010, the 106th Tennessee General Assembly passed a comprehensive set of education reforms with few amendments on January 15, which were then officially signed into law by Governor Bill Haslam the very next day under the First to the Top (FTTT) Act. Only eleven out of out of 114 state legislators across both legislative chambers voted not to pass the bill (one of which was Republican), thus signifying the bipartisan ethos undergirding the advancement of educator evaluation reform throughout the state (Tennessee General Assembly, n.d.).

While the expeditious legislative process surrounding the passage of the FTTT Act appeared seemingly straightforward, the same was certainly not true for the coalition-building required to gain stakeholder buy-in. The Tennessee Department of Education (TDOE) sagely facilitated collaboration among several stakeholders early on in the RTTT application process, long before a bill ever reached the house and senate floor for consideration. Support from the Tennessee Education Association (TEA), the state teachers union, was quite possibly the most vital, as it was cited as instrumental during the state's bid to land RTTT funds (U.S. Department of Education, 2010b). However, the devil was in the details of the law. After teacher evaluation reform was codified and implementation was underway, TEA began to renounce aspects of the reformed system, specifically its use of value-added data to evaluate teachers, which became the subject of litigation in later years.

In addition to laying out the reformed educator evaluation system in broad strokes, the FTTT Act commissioned a teacher evaluation advisory committee to develop and recommend more detailed guidelines and criteria for annual evaluation. As per provisions within the FTTT Act, the teacher evaluation advisory committee comprised 15 members representing a diverse set of stakeholders, including the state education commissioner, serving as committee chair; the executive director of the state board of education; a member from each of the education committees for the state house and senate; two K-12 public school teachers separately appointed by the senate and house; and nine governor appointees. Governor appointments to the committee were to include three additional K–12 public school teachers, two public school principals, one director of a school district, and one parent of a currently enrolled public school student—with the remaining representing other stakeholder interests. The teacher evaluation committee was formed within 30 days of the bills effective date and was to present evaluation guidelines and criteria for adoption effective July 1, 2011.

In the end, the teacher evaluation committee met the timeline specified in the FTTT Act (Loewus, 2011). The new teacher evaluation system was in full effect throughout the state at the start of the 2011–12 academic year. By November of 2011, not long after its kickoff, Governor Bill Haslam publicly proclaimed that Tennessee was the "the focal point of the education reform in the nation" (Winerip, 2011) a remark surely made in the same spirit as the state's new motto for education: "First to the Top."

7.3 Theory and Design Elements of the Reformed Teacher Evaluation System

The theory of change underlying Tennessee's reformed teacher evaluation system as a tool to improve quality within the teacher workforce rests on two primary functions. First, formative evaluations involving feedback mechanisms assist with the development of teacher productivity in alignment with a school's mission (i.e., to generate growth in student learning). Second, summative evaluations provide school leaders with information to assess teacher performance to facilitate staffing decisions. These two functions of evaluation not only reflect the tenets described by the personnel and human resource management research literatures (e.g., Arnold, 2005; Gomer-Mejia et al., 2015) but also reflect the legislative language and policy directives at the federal and state levels describing the intended design and implementation of rigorous evaluations systems for teachers (e.g., Tenn. Code Ann. $ 49-5-501-515, 2012; U.S. Department of Education, 2009a). In fact, RTTT grant guidelines explicitly called for evaluation systems capable of "developing teachers …, including by providing relevant coaching, induction support, and/or professional development" and "removing ineffective tenured and untenured teachers … after they have ample opportunities to improve …" (U.S. Department of Education, 2009a, p. 9).

In July 2011, the Tennessee State Board of Education (TSBE) approved the Tennessee Educator Acceleration Model (TEAM) as the new default classroom observation evaluation model across the state. Along with TEAM, the board approved three alternate classroom observation evaluation model options for districts with demonstrated satisfactory performance: Project Coach (COACH), Teacher Effectiveness Measure (TEM), and Teacher Instructional Growth for Effectiveness and Results (TIGER). A fifth model, the Achievement Framework for Excellent Teaching (AFET), was later approved and first implemented in the 2012–13 school year. By and large, all five classroom observation evaluation models follow a similar structure and primarily differ in the specifics of their observation rubric domains and the duration and frequency within which classroom observations are conducted. Thus, the focus of the underlying theory and design elements described below are largely in reference to the default classroom observation evaluation model—TEAM. The descriptions provided are largely applicable to the alternate classroom observation evaluation models; however, notable distinctions are elaborated as necessary.[2]

[2] All alternate evaluation models satisfy overarching guidelines established by TSBE, including (1) calculation of evaluation ratings based on the fifty percent (50%) quantitative data, including

7.3.1 Design Elements of TEAM

In accordance with RTTT federal guidance, TEAM incorporated three main characteristics to evaluate teachers: (1) annual classroom observation and evaluation of teachers; (2) inclusion of multiple categories on which to evaluate teacher effectiveness, with student growth accounting for a significant portion of the overall measure of effectiveness; and (3) use of evaluation ratings to inform decisions regarding professional development, compensation, promotion, retention, and tenure. Although Tennessee ultimately approved four alternate classroom observation models for districts to use, each model incorporates all four of the aforementioned criteria.[3]

7.3.1.1 Annual Classroom Observations

A primary purpose of evaluation is to provide teachers with instructional feedback; to that end, classroom observations are pivotal. Depending on the type of evaluation model (e.g., TEAM, TEM, Project COACH), the number of observations a teacher receives can vary from four to ten per year. Although the specific structure of feedback mechanism depends on the evaluation model (e.g., duration and frequency of announced versus unannounced observations, observation rubric domains), all teachers participate in a post-observation feedback conference with the observer and receive written feedback intended to provide certain forms of reinforcement (an area of strength) and refinement (an area in need of improvement).

Under TEAM, “observation pacing,” or the number of observations within a year that an educator must receive, is determined by a combination of the educator’s licensure status and previous year’s evaluation rating, as depicted in Table 7.1. At least half of all observations must be unannounced. In cases where a teacher did not receive an evaluation rating in the previous year (e.g., teacher is new to the profession or returning from maternity leave), the maximum number of observations based on licensure status is then required. Interim teachers who have not completed a minimum of 120 days of service at their school, whether due to having taken extended leave, transferred to a different school mid-year, or transitioned to another role, are granted partial year exemption (PYE) status, which would preclude or delay their eligibility for certain forms of benefits such as tenure, bonuses, or salary increases.

student achievement and growth measures; (2) completion of observations by certified evaluators; (3) reliance on a research-based observation rubric that addresses the four domains of planning, environment, professionalism, and Instruction; and (4) use of personal conferences to discuss strengths, weaknesses and remediation, and classroom observation visits.

[3] In addition to providing an evaluative framework for teachers, TEAM also established a corresponding framework for the evaluation of school leaders. School leaders are similarly assessed based on a combination of observation, input from school staff, and student data; however, the evaluation of school leaders relies on an entirely separate set of standards called the Tennessee Instructional Leadership Standards (TILS), for which a more detailed discussion falls outside of the scope of this current chapter.

Table 7.1 Standard TEAM observation pacing

Educator licensure status	Previous individual growth or level of overall effectiveness (LOE)[a]	Minimum required observations	Minimum required observations per domain	Minimum number of minutes per school year (min)
Practitioner[b]	Levels 1–4	Six (6) domains observed with a minimum of three (3) domains observed in each semester	3 instruction 2 planning 2 environment	90
	Level 5	One (1) formal observation covering all domains first semester; two (2) walkthroughs second semester	1 instruction 1 planning 1 environment	60
Professional	Level 1	Six (6) domains observed with a minimum of three (3) domains observed in each semester	3 instruction 2 planning 2 environment	90
	Levels 2–4	Four (4) domains observed with a minimum of two (2) domains observed in each semester	2 instruction 1 planning 1 environment	60
	Level 5	One (1) formal observation covering all domains first semester; two (2) walkthroughs second semester	1 instruction 1 planning 1 environment	60

Note Adapted from "Observation Guidelines" document made available by the Tennessee Department of Education. A district using the TEAM model may choose to allow observers to combine domains during classroom observations provided the requisite minimum time, semester, distribution, and notice (announced vs. unannounced) are met. [a] Districts may elect to base pacing on a teacher's previous year individual growth or on level of overall effectiveness pursuant to local policy. [b] The practitioner status applies to all other non-professional license types such as adjunct, international, and initial licenses, including the apprentice license

TEAM evaluates all teachers based on a standard general educator rubric adapted from the Charlotte Danielson Framework for Teaching covering the three domains of "instruction," "planning," and "environment." As illustrated in Fig. 7.1, the three domains contain various sub-domains for which teachers must be rated throughout an academic year.[4] Teachers must receive at least one observation focused on each of the three domains each year, though districts have discretion to allow for the simultaneous observation of the instruction domain along with either the planning or environment domain during the same classroom visit. After each observation, evaluators are expected arrange a post-observation conference with the teacher to deliver feedback in an area of refinement and reinforcement and grant an opportunity for the teacher to engage in self-reflection on their practice.

Aside from the observations and feedback associated with the three main observation domains, there are other forms of observation and feedback incorporated into TEAM. Upon completion of the academic testing season, evaluators also score teachers in an additional "professionalism" domain, which reflects evidence of a teacher's professional growth and learning, use of data, school and community involvement, and leadership. Additionally, evaluators conduct 10–15 min unscored classroom "walkthroughs" to provide rapid, narrowly focused feedback to teachers who receive fewer classroom observations after having previously scored highly effective. Finally, in addition to the post-observation conference, evaluators are expected to arrange summative conferences with educators to holistically review their evaluation results throughout the academic year and discuss areas for continued refinement.

All classroom observations are conducted by trained, certified observers. TDOE regularly conducts certification and recertification training and requires observers whose ratings systematically and drastically differ from student growth score ratings for the same group of teachers to participate in additional training. According to first-year implementation reports, principals and assistant principals conducted the vast majority of classroom observations for teachers, while instructional coaches, lead teachers, department heads, and external certified observers conducted a small share of observations (Pepper et al., 2012).

Notably, observation scores are only partially greviable under TEAM, meaning that in teachers cannot formally submit a complaint about their observation score unless a violation of standard evaluation policy arises.

7.3.1.2 Student Achievement Measures

Under TEAM, teachers are partly assessed on the level of student achievement. The selection of an achievement measure is individualized and flexible from teacher to teacher, as each teacher and his or her evaluator mutually decide on a test-based or

[4] Additional rubric guidance is provided specifically for early childhood and pre-K educators; alternative, gifted, and special education teachers; English as a second language educators; and physical education teachers (Tennessee Department of Education, n.d.-a).

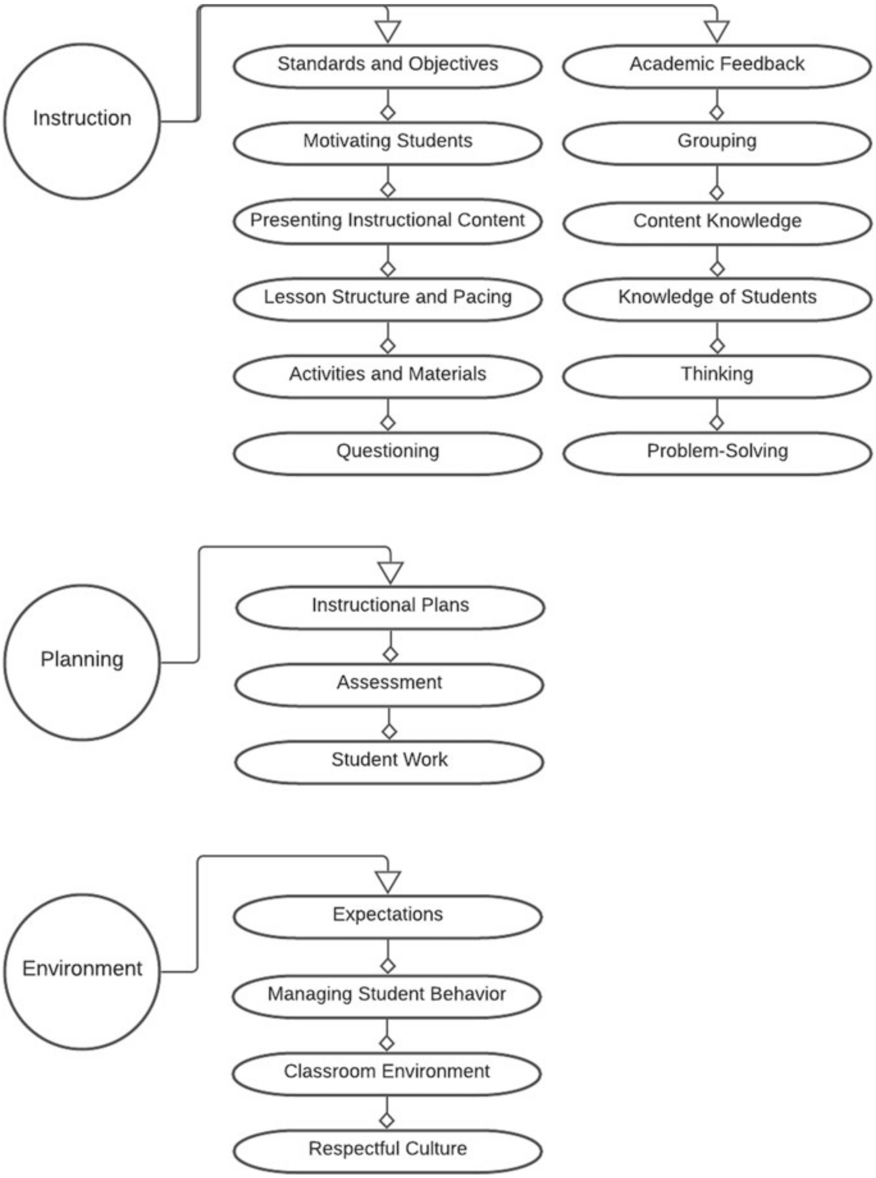


Fig. 7.1 TEAM classroom observation rubric domains

alternative measure of student achievement (Tennessee Department of Education, 2020a). For instance, an elementary or middle-grade teacher grade could select a subject-specific state assessment or could equally select an off-the-shelf assessment such as *STAR Math* or *Reading Recovery* as the measure of student achievement feeding into their LOE score. Likewise, high school teachers could select among from subject-specific end-of-course state assessments, graduation rate, ACT/SAT assessments, or AP/IB assessments as their achievement measure. Each year, TDOE provides a list of acceptable achievement measures specific to particular teaching areas and schooling levels.

7.3.1.3 Test Score Growth and Estimated Teacher Value-Added

In addition to the *level* of student achievement, teachers are evaluated based on *growth* in student achievement, where growth is measured through value-added methods. Tennessee's adoption of value-added (i.e., TVAAS) predated the comprehensive reform of its evaluation system for teachers by several years. Developed by statistician William Sanders and his associates, TVAAS was first put in place as a teacher evaluation tool for school programs in Tennessee beginning in 1993 (Braun, 2005; Kupermintz, 2003). TVAAS measures teacher effectiveness on the basis of student gains and hinges on a multi-level calculation that blends the estimation of the average performance gains in each school system, for each year, grade, and academic subject, and the average performance of each teacher's students, relative to the system's performance (Ballou et al., 2004; Kupermintz, 2003). For details of the TVAAS methodology and an example of the estimation of system, school, and teacher effects, see Sanders et al. (1997).

Naturally, the viability of inferences drawn from TVAAS estimates hinges on the extent to which they adequately capture teachers' unique contributions to student learning. The literature on this topic is fairly robust (e.g., Ballou & Springer, 2015; Ballou et al., 2004; Kupermintz, 2003; Sanders & Horn, 1998), and it would be out of this chapter's scope to fully review its nuances. However, a few points warrant discussion to summarize the general takeaway. First, a chief critique raised against TVAAS is that its methodology does not explicitly control for student background factors, which doubtlessly may influence a student's initial levels of achievement and achievement gains. While modified TVAAS models that include controls for observable student characteristics commonly used in other value-added estimation approaches (e.g., socioeconomic status and demographics) have shown the exclusion of such controls have negligible impact on estimation of teacher effects, inclusion of a simple fixed effects estimator does not (Ballou et al., 2004), which raises questions around whether TVAAS fully adjusts for the influence of student background in the estimation of teacher effects. A second concern pertains to the accuracy of TVAAS estimates. In TVAAS, the accuracy of estimated teacher effects depends on the amount of data available for each teacher—estimates for teachers with less data (i.e., less students taught in a particular year) are less accurate than those of teachers with more data. Furthermore, teacher effects are "shrunken" toward the system's average—when student data are scarce, a teacher is assumed by the model to perform at the average level of his or her school system. The fewer the students, the stronger the pull toward the overall system mean, which is most concerning for teachers working in low-performing school districts with fewer student test data.

7.3.1.4 Calculation of Teachers' Level of Overall Effectiveness

Notwithstanding concerns surrounding the validity and reliability of TVAAS, it nevertheless continues to function as a central performance metric Tennessee's reformed evaluation system, though it is one of many considering the evaluation system's multiple-measures approach. In fact, TVAAS necessitates the incorporation of alternative measures of teacher performance to rightfully buttress against its weaknesses.

Across all evaluation models, a teacher's summative Level of Overall Effectiveness (LOE) is evaluated across three separate dimensions: (1) classroom observations, (2) student achievement, and (3) student growth or TVAAS (Tennessee Department of Education, 2019a, n.d.-b).[5] At the time of this writing, teachers for whom standardized tests are applicable and available (i.e., "tested teachers") receive individual TVAAS growth data that comprises 35% of their LOE. Teachers for whom standardized tests are not applicable (i.e., "non-tested teachers"), this individual growth measure is replaced by a school-wide measure of student growth, and only contributes 15% of their LOE. Student achievement is weighted at 15% of a teacher's LOE score, regardless of whether they are a tested or non-tested teacher. The remaining 50% of the evaluation score is calculated from classroom observations.

The three key performance metrics (student growth, student achievement, and observations) combine to create the comprehensive measure of teacher effectiveness or the LOE. The LOE is calculated as a continuous scale that ranges from 0 to 500 and is cut into five discrete categories that are used as the primary evaluation ratings. Teachers with an LOE below 200 receive a Level 1 rating, indicating "Significantly Below Expectation"; between 200 and 275 are categorized into Level 2, "Below Expectation"; between 275 and 350 are Level 3, "At Expectation"; between 350 and 425 are Level 4, "Above Expectation"; and teachers with an LOE of 425 and above are categorized into Level 5, "Significantly Above Expectation."

[5] While the use of student test score data is required as part of formal personnel evaluation, initially, the use of three years of TVAAS for student test score growth was required to measure teacher effectiveness. The calculation of the TVAAS composite measuring student growth attributable to teacher effectiveness has changed overtime to accommodate implementation challenges with the administration of standardized tests, the specifics of which will be discussed in more detail below.

7.3.2 Evaluation as a Conduit for Human Capital Management

Aside from identifying levels of teacher quality for the purposes of teacher development, evaluation ratings are incorporated into decisions around compensation, promotion, retention, and tenure. A number of notable statewide and districtwide policies have been tied to teacher evaluation throughout the years, specifically regarding a teacher's eligibility for tenure and alternative compensation systems for teachers.

7.3.2.1 Reforms to Teacher Tenure Eligibility

Beginning in April of 2011, and virtually in tandem with reforms to the evaluation system, Tennessee legislated and implemented comprehensive changes to tenure protections for teachers, requiring that tenure decisions to be linked to evaluation ratings. Under Tennessee's prior tenure statute, teachers were eligible to receive tenure conditional on serving a three-year probationary period in the same school district, meaning that tenure decisions did not take into account measures of teacher performance or effectiveness. The 2011 tenure statute made several amendments to the tenure eligibility process. First, the probationary period was extended from three years to five years. Second, teachers without tenure prior to the passage of the law were required to receive a LOE rating that placed them in one of two highest performance categories under reformed evaluation system ("Above Expectation" or "Significantly Above Expectation") during the final two years of the extended five-year probation period in order to become eligible for tenure. Teachers who did not receive tenure status at the end of their five-year probation period were either rehired under a year-to-year contract or prone to being dismissed. Finally, the new statute mandated that tenured teachers continue to demonstrate high levels of performance. If a tenured teacher's LOE dropped below Level 3 for two consecutive years, the teacher would lose tenure protections and must cycle through the entire process again to regain tenure status. However, teachers who received tenure under the old statute (before 2011) were grandfathered in, such that the maintenance of their tenure status was not at all contingent on evaluation ratings (Tenn. Code Ann. § 49-5-501-515, 2012).

7.3.2.2 Alternative Compensation Systems for Teachers

There are a number of prominent examples of alternative compensation systems tied to teacher performance under the reformed evaluation system, all of which were exclusively administered to teachers to specific districts or schools, either because the systems remained within the piloting stage, were tied to district-specific funding sources, were designed to improve the pool of teaching candidates in hard-to-staff schools, or a combination thereof. Beginning in the 2011–12 school year, three separate initiatives were launched to support the implementation of strategic teacher compensation plans in the Tennessee public schooling system: The Competitive Supplemental Fund (CSF), the Innovation Acceleration Fund (IAF), and the Tennessee Teacher Incentive Fund (TN TIF). All three initiatives were designed to support district efforts to implement alternative means to compensate teachers that differed from the standard statewide Minimum Salary Schedule, which pays teachers based on highest degree earned and teaching experience. CSF, IAF, and TN TIF targeted about $30 million of funding provided by the US Department of Education's Teacher Incentive Fund, the federal Race to the Top grant, as well as several private foundations to 14 districts serving almost 200 schools across the state over the duration of five years (Ballou et al., 2016). The implemented compensation plans varied across the 14 districts, but generally provided performance bonuses to highly effective teachers as well as extra pay for professional development and leadership activities. Separately, in the spring of 2013, the TDOE piloted a pair of bonus programs designed to attract and retained high-performing teachers in low-performing school settings. In that year, teachers receiving the highest LOE rating (Level 5 status) were eligible to receive a $7000 signing bonus if they voluntarily transferred to teach in a Priority School, the state's official designation for the bottom 5% of lowest performing schools based on a composite proficiency rate (success rate) for all students in a school. Similarly, Level 5 teachers already teaching in a Priority School were eligible to receive a $5000 retention bonus if they chose to remain teaching in a Priority School for an additional school year.

7.4 Teacher Evaluation in Practice: Reports on the Consequential Validity of Reform

Tennessee's reformed teacher evaluation system integrated many of the design elements promoted in RTTT and as such was encouragingly well-equipped from a theoretical basis to support the development and enhancement of the teaching workforce throughout the state. Yet a number of questions naturally arise as to how theory and design translated into practice. How did the reformed evaluation system function? How did major stakeholders most directly affected by the system—teachers and school leaders—perceive the changes made to evaluation? How did they utilize it to improve their practice? And, finally, what challenges arose to affect the system's implementation?

From a general standpoint, the implementation of teacher evaluation reform rolled out as scheduled. Each year since 2011, over sixty-thousand public school teachers have been evaluated across Tennessee. For example, in the first year of implementation, only a minority of teachers received the lowest two LOE ratings (7%), as shown

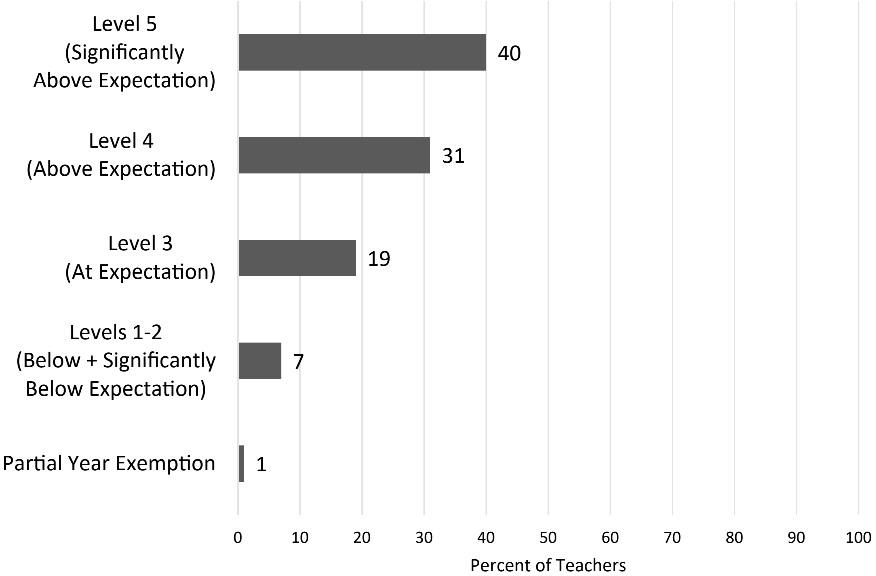


Fig. 7.2 Distribution of teachers across level of effectiveness (LOE) ratings, 2011–12 academic year. *Note* Information retrieved from Table 7.1 presented in Koedel et al. (2017)

in Fig. 7.2. The remainder of teachers who received ratings performed "At Expectation" (19%), "Above Expectation" (31%), or "Significantly Above Expectation" (40%). The skewed nature of the distribution of LOE rates has remained a persistent problem under TEAM and has become more stark in recent years.

Fortunately, the implementation of teacher evaluation reform proceeded in tandem with deliberate efforts to monitor its progress, largely in part due to the annual joint effort by TDOE and the Tennessee Education Research Alliance (TERA) to gather information about schools across the state through the administration of the *First to the Top Surveys* to teachers, administrators, and certified staff from 2011 to 2014 and subsequent *Tennessee Educator Surveys* from 2015 and onward. The annual surveys, as well as other studies conducted by independent investigators, have provided a wealth of information on the reformed teacher evaluation system, namely how teachers and administrators perceived and experienced changes to the teacher evaluation process.

7.4.1 Perceptions of Teacher Evaluation Reform

According to reports generated from the *First to the Top Surveys* and *Tennessee Educator Surveys,* teachers tend to increasingly perceive the reformed evaluation system as capable of improving their teaching and student learning. In 2019, 76%

of teachers reported that they agreed evaluation improved their teaching, up from 38% of teachers in 2012 (Tennessee Department of Education, 2019b; Pepper et al., 2012).

However, a number of sources highlighted particular areas of concerns, both among teachers and school administrators largely tasked with carrying out the classroom observations required of the teacher evaluation system. Survey reports from TDOE indicate that teachers valued forms of feedback received outside of the evaluation system to a greater extent than forms of feedback received through the evaluation process directly (Tennessee Department of Education, 2016), which has highlighted the need to improve the quality and utility of feedback delivered to teachers in both formative and summative feedback conferences. Furthermore, in 2018, despite an increased perceived benefits generated from evaluation to both teaching and learning, 50% of teachers reported the evaluation process posed a significant burden with a number of teachers commenting that time, resource constraints, and anxiety associated with evaluation being major causes of experienced burden (Tennessee Department of Education, 2018).

External reports highlight similar concerns among administrators, specifically regarding the large amount of time needed to complete the evaluation process (Campbell & Derrington, 2019; Olson, 2018). Prior to teacher evaluation reform in 2011, principals were previously been required to evaluate teachers only once every five years (Olson, 2018), which clearly fell far below the now required annual observations that could frequent up to six times per year depending on the teacher's licensure status and prior rated performance. Likely stemming from these expressed constraints, survey reports have found some evidence of the evaluation system having negatively affected relationships between teachers and principals (Derrington & Martinez, 2019).

7.4.2 Utilization of the Teacher Evaluation System

One of the main purposes of a robust evaluation process is to provide useful feedback that teachers may use to improve their instructional practice. Reports from first-year implementation surveys indicated that teachers were largely being observed the expected number of times and for the expected duration as per their district's selected classroom observation evaluation model (Pepper et al., 2012). Teacher also reported a vast distribution in the time spent reviewing post-observation feedback with their evaluator, though some models required more frequent and shorter observations and feedback sessions by design (i.e., COACH). At least 50% of teachers reported spending at least 10 min reviewing feedback from short observations and 30 min reviewing feedback from lesson-length observations, with the majority of post-observation conferences taking place in a timely fashion within 10 days of a teacher being observed (Pepper et al., 2012).

Although the evaluation system offers multiple metrics of teacher performance (i.e., observation ratings, achievement measures, value-added or growth scores, the summative LOE rating), a number of reports have indicated that teachers and administrators tend to rely on only a subset of those measures more than others (Campbell & Derrington, 2019; Goldring et al., 2015; Pepper et al., 2012). More specifically, principals have expressed low confidence in the test score-based components of teacher evaluation ratings (Campbell & Derrington, 2019) and favor teachers' observation ratings when making human capital decisions, largely citing the perceived consistency, transparency, and specificity provided by observation data (Goldring et al., 2015).

Utilizing evaluation results to improve teacher practice, however, was a stark challenge in the early years of the evaluation system's implementation. Based on first-year implementation reports, less than 5% of teachers strongly agreed that the feedback they received from being observed informed their professional activities (Pepper et al., 2012). Later analyses also found no evidence that teachers alter their time investments in professional improvement or adjust professional improvement activities based on their summative evaluation ratings (Koedel et al., 2019). Researchers cite insufficient incentives to encourage teachers to respond to their ratings as a potential explanation for the disconnect between evaluation results and professional development activities among teachers (Koedel et al., 2019). Aside from the pretenure process, which is largely confined to early-career teachers, and performance-based bonuses in select districts, there are few mechanisms imbedded in the teacher evaluation system to explicitly incentivize improvement in teacher evaluation ratings.

In attempt to connect evaluation with meaningful development opportunities, TDOE has developed various supports to help teachers further improve. One example includes the piloted teacher-pairing program called the Instructional Partnership Initiative (IPI), which paired higher- and lower-performing teachers based on areas of strength and growth in specific observation domains. With guidance from their principals, paired IPI teachers collaborated to improve their skills in the areas identified as in need of improvement. In addition, TDOE has integrated a micro-credentialing program wherein teachers attain professional development points (PDPs) for having completed personalized, competency-based professional development activities, which they may accrue for licensure advancement or renewal. To earn PDPs, professional activities must center around developing content or pedagogical knowledge, enhancing educator effectiveness in a specialized practice area (e.g., English as a second language, data utilization), or developing competency in student social and emotional health and well-being (Tennessee State Board of Education, 2020a).

7.5 Implementation Challenges to Teacher Evaluation Reform

Unanticipated challenges may sometimes arise to affect the quality of implementation of a policy reform. Tennessee's teacher evaluation reform has been no exception. Two kinds of issues considerably affected teacher evaluation in Tennessee since its reform in 2011—judicial challenges against aspects of the reformed system and problems with testing administration under a new state standardized testing regime.

7.5.1 Judicial Challenges

Three seminal legal challenges were raised against one of the centerpiece aspects of the reformed teacher evaluation system in Tennessee—the use of value-added as a measure of teacher effectiveness. The TEA filed two lawsuits in 2014 challenging the statistical methodology underlying TVAAS and its use to determine teacher effectiveness: *Trout* v. *Knox County Board of Education* and *Taylor v. Haslam*. Both cases were filed, eventually combined as one under the *Trout* suit, on behalf of two teachers who were denied a bonus due to low TVAAS ratings that were based on only a subset of the teachers' students. The *Trout* case involved the erroneous exclusion of the teacher's students from their TVAAS calculation, while the *Taylor* case involved the exclusion of advanced students taught by the teacher who completed local assessments in lieu of the standardized test. The plaintiffs' argued that the use of TVAAS estimates to evaluate teachers was arbitrary and capricious and, therefore, could not pass rational scrutiny. However, the court eventually ruled against their favor, stating that "While it may be a blunt tool, a rational policymaker could conclude that TVAAS is 'capable of measuring some marginal impact that teachers can have on their own students'" (*Taylor v. Knox County Board of Education*, 2016). Ultimately, the court acknowledged the concerns about the statistical imprecision of TVAAS but concluded that the judiciary was not empowered to rule out its use altogether.

Later in 2015, the TEA filed another federal lawsuit on behalf of a group of teachers, *Wagner v. Haslam*, but specifically challenging the use a school-wide growth measure to evaluate teachers in non-tested subjects such as the arts or physical education. The court, echoing the earlier decision established by *Trout*, rejected the teachers' arguments. While concluding that the use of TVAAS was constitutional, the court decision did acknowledge the "unfair results" for certain teachers despite not rising to the level of being irrational (Wagner v. Haslam, 2015).

7.5.2 Testing Administration Issues

The reformed teacher evaluation system in Tennessee has been affected by a history of testing administration issues, particularly since the state's transition to a new standardized test. When teacher evaluation reform was first passed and implemented in 2011, Tennessee had been working for several years toward aligning its standards with Common Core through a multi-state consortia known as Partnership for Assessment of Readiness for College and Careers (PARCC). However, due to backlash over Common Core, in 2014, and six months prior to the start of testing, the Tennessee legislature voted to pull out abruptly from PARCC, marking the state's move away from Common Core standards altogether. Shortly thereafter, the state identified a vendor to develop a new test, TNReady, to start pilot testing in the fall of 2015. TNReady was a two-part test designed to be fully administered in an online format.

In 2016, the first year that TNReady was to be fully administered, software issues prevented students from being able to log in successfully (Gonzales, 2016). The online test was canceled entirely, while only high school students were able to complete a paper version of the test as the vendor was unable to deliver paper tests on time to elementary and middle schools, citing the high number of tests needed which fell a little under 10 million (Gonzales, 2016, 2018). TDOE eventually fired and replaced its vendor, scheduling for a replacement vendor to begin testing administration.

In 2018, TNReady was plagued with various login and disruption issues yet again, forcing the state to extend the online testing window. Even more worrisome, the vendor incorrectly graded completed paper tests in the previous year. Due to its checkered history with administration issues, educators and lawmakers expressed low confidence and morale with TNReady and cited frustration after having committed considerable resources to prepare for the test (Aldrich, 2018; Gonzales, 2018). The state responded by convening an assessment task force to ensure problems did not persist into the subsequent testing year. The task force's audit ultimately concluded that testing issues arose due to a number of reasons, including TDOE oversight of its work plan with the testing vendor and an unauthorized change made by the vendor that resulted in login issues (Tennessee Comptroller of the Treasury, 2018).

In both 2016 and 2018, TDOE made a number of accommodations in the light of the administration issues with TNReady. First, the impact of TNReady results on teacher evaluation was phased in such that TNReady could account for only 10% of teachers' LOE score for teachers in tested subjects and grades—the remaining 25% would come from previous years of testing results from the former standardized test whenever available (McQueen, 2015). Furthermore, teachers could opt to have TNReady growth scores contribute the entirety of the 35% growth measure if doing so would benefit a teacher by yielding a higher LOE score (McQueen, 2015). During the 2017–18 academic year, teachers with TNReady data also had the option to nullify their entire LOE generated for the school year at their discretion (Tennessee Department of Education, n.d.-c). If nullified, the use of LOE scores was prohibited when making decisions regarding employment termination and compensation. Subsequently, TDOE reduced the use of student growth data from 35 to 15% for non-tested teachers (Tennessee Department of Education, n.d.-b).

Tennessee has been forced to face test administration issues yet again, but this time through no fault of any vendor but due to the presently ongoing COVID-19 pandemic. In response to the pandemic, the state board of education passed a series of emergency rules to waive the calculation of TVAAS and LOE scores for teachers entirely during the 2019–20 academic year (Tennessee Department of Education, 2020b).

7.6 Effects of Teacher Evaluation Reform

Now that teacher evaluation reform has undoubtedly taken root in Tennessee over the past several years; it is worth taking stock of the available evidence regarding its impact. Two notable areas of study elucidate the extent to which teacher evaluation reform has benefited teachers and, in turn, students—these areas attend to the system's impact on, firstly, teacher performance and student achievement and, secondly, its impact on teacher retention.

7.6.1 Teacher Performance and Student Achievement

A primary aim of the teacher evaluation system is to improve teaching and student performance. While select studies that have shown specific programs imbedded within Tennessee's teacher evaluation system have caused improvements in student achievement, including the IPI program (Papay et al., 2020) and the priority school teacher retention bonus program (Swain et al., 2019), estimating the comprehensive impact of teacher evaluation reform on teacher performance and student achievement is quite tricky. Unlike other states that piloted their reforms in a subset of districts, Tennessee chose to implement its reforms statewide simultaneously. Being that all teachers were subject to the new evaluation process, identification of a valid comparison group is infeasible.

Nevertheless, research affiliates of TERA have—to the best of their ability—investigated the extent to which teacher evaluation reform and the professional development initiatives married to the system have been associated with improvements in teacher performance. Using a wide array of performance measures, including student test scores in mathematics, English language arts, science, and social studies; classroom observation scores; and TVAAS, a recent analysis examined the average rate of improvement among teachers in Tennessee pre- and post-implementation of teacher evaluation reform. The results of the analysis revealed teachers have improved at a much steeper rate in the years subsequent to the implementation of the reformed evaluation system, from 2013 to 2015, than the rate of improvement just prior to reform, between 2008 and 2010 (Papay & Laski, 2018). Based on the currently available evidence, it is likely the evaluation system itself as well as concurrent initiatives to facilitate teacher development were the main factors producing the observed improvements in teaching.

7.6.2 Teacher Retention

Also of interest is the system's impact on teacher retention, especially as it was in part intended to inform human capital decisions, including the hiring, assignment, and dismissal of teachers within, across, and from schools. Research based on time series analyses suggests the rollout of a statewide evaluation system was associated with increased turnover among teachers; however, there was comparably greater retention of more effective teachers in the years following the system's reform (Rodriguez et al., 2020). Moreover, differences in turnover between highly and minimally effective teachers were most apparent in urban school districts and low-performing schools (Rodriguez et al., 2020). Such increases in turnover are perhaps unsurprising considering teacher reports characterizing the evaluation process as burdensome due to stress and time and resource constraints arising from undergoing more extensive observation and consultation with evaluators. It is also plausible that turnover rates were comparably lower among more effective teachers, especially in the light of prior research suggesting that higher ratings under the evaluation system improved teachers' job satisfaction and perceptions of their work environment relative to lower ratings (Koedel et al., 2017).

While the evidence of teacher evaluation reform's impact on teacher retention in Tennessee is not fully encouraging, there are silver linings. Specifically, prior research documents positive effects on teacher retention associated with the piloted retention bonus program imbedded within the teacher evaluation system. Teachers eligible to receive a $5000 bonus upon being rated "Significantly Above Expectations" were more likely to remain teaching in low-performing school settings, specifically when teaching tested subjects and grades (Springer et al., 2016). Such findings demonstrate the evaluation system's ability to facilitate positive and equity-oriented teacher retention patterns, particularly when it is equipped with targeted and strategic incentives.

7.7 Lessons Learned from a System with Promise

Over the past decade within the United States, teacher evaluation reform has represented a prevalent strategy with promise to strengthen the teaching profession and improve student learning. Some evidence supports the positive impact teacher evaluation reform has had in Tennessee for both teachers and students; however, the system has not been without faults and challenges in both its design and implementation. In spite of this, how teacher evaluation reform came about in Tennessee was nothing short of remarkable and offers several key lessons that are applicable not only for school systems considering changes to teacher evaluation but, at least a few of which, are generally transferrable to school systems considering other forms of widescale policy change.

On the matter of policy formulation and adoption, evaluation reform—as is the case with any major policy reform—necessitates a complex coalition of stakeholders that are able and willing to coalesce around a singular aim. As shown by the exchange of resources between the federal government and state education agencies supporting teacher evaluation reform in Tennessee, the bipartisanship support undergirding the legislative passage of teacher evaluation reform in the state legislature, and the partnership between TDOE and the TEA when designing the evaluation system, building such a coalition is a delicate and time-consuming endeavor. Large-scale, impactful policy change is not the sole responsibility of local actors, as demonstrated by the competitive federal grant program that largely prompted evaluation reform to take place in Tennessee in the first place. Nor does policy change arise from the top-down from a centralized entity, as the reform process in Tennessee involved the input of local policymakers and practitioners when establishing the system's design. Yet the coalition advancing evaluation reform in Tennessee was not a static arrangement, stakeholders that initially worked in partnership to bring out a reformed evaluation system later advocated for divergent interests and aims that were eventually settled in court.

Tennessee's reformed teacher evaluation system also offers lessons in the area of implementation. Unlike other states implementing reformed teacher evaluation systems post-RTTT, Tennessee chose to roll out a comprehensive reformed system statewide within the timespan of one calendar year. While an impressive feat, the rapid rollout of the system placed intense constraints on many stakeholders, especially teachers and school administrators. A more measured rollout, perhaps in cohorts organized by school districts, may have proceeded more seamlessly. Nevertheless, one could characterize the implementation of Tennessee's teacher evaluation system as flexible and nimble when circumstances necessitated. Fortunately, when faced with test administration issues and the COVID-19 pandemic, TDOE and state policymakers were amenable to departing from rigid policy guidelines and requirements initially established under the evaluation system.

Tennessee's teacher evaluation system and the process by which it has been implemented embodies elements of both feedback and accountability. TDOE, TERA, and other agencies and independent stakeholders have continually monitored the progress of the system, often in service of identifying areas for continuous improvement based on data and experience. And while the heart of the teacher evaluation system privileges mechanisms for direct feedback, it has—to varying degrees—been coupled with incentives and supports for teacher improvement and retention of effective teachers. Moving forward, researchers, policymakers, and practitioners should attend to ways to improve systems, not only comprehensively but incrementally as well. Given time and continued investment, sweeping policy reforms can bring about meaningful educational change.

References

Aldrich, M. W. (2018, August 6). Declaring "no confidence" in TNReady, Memphis and Nashville superintendents call for pause in state testing. *Chalkbeat: Tennessee*. Retrieved from https://tn.chalkbeat.org/2018/8/6/21105501/declaring-no-confidence-in-tnready-memphis-and-nashville-superintendents-call-for-pause-in-state-tes

Arnold, E. (2005). Managing human resources to improve employee retention. *The Health Care Manager, 24*(2), 132–140.

Ballou, D., Canon, K., Ehlert, M., Wu, W. W., Doan, S., Taylor, L., & Springer, M. (2016). *Final evaluation report: Tennessee's strategic compensation programs: Findings on implementation and impact 2010–2016.* Retrieved from https://peabody.vanderbilt.edu/TERA/files/TIF_FINAL_7.1.16.pdf

Ballou, D., Sanders, W., & Wright, P. (2004). Controlling for student background in value-added assessment of teachers. *Journal of Educational and Behavioral Statistics, 29*(1), 37–65.

Ballou, D., & Springer, M. G. (2015). Using student test scores to measure teacher performance: Some problems in the design and implementation of evaluation systems. *Educational Researcher, 44*(2), 77–86.

Braun, H. I. (2005). *Using student progress to evaluate teachers: A primer on value-added models.* Princeton, NJ: Policy Information Center, Educational Testing Service. Retrieved from https://www.ets.org/Media/Research/pdf/PICVAM.pdf

Campbell, J. W., & Derrington, M. L. (2019). Principals' perceptions of teacher evaluation reform from structural and human resource perspectives. *Journal of Educational Supervision, 2*(1), 58–77.

Derrington, M. L., & Martinez, J. A. (2019). Exploring teachers' evaluation perceptions: A snapshot. *NASSP Bulletin, 103*(1), 32–50.

Glander, M. (2017*). Selected statistics from the public elementary and secondary education universe: School year 2015–16* (NCES 2018-052) [Table 2]. U.S. Department of Education. Washington, DC: National Center for Education Statistics. Retrieved from http://nces.ed.gov/pubsearch

Goldring, E., Grissom, J. A., Rubin, M., Neumerski, C. M., Cannata, M., Drake, T., & Schuermann, P. (2015). Make room value added: Principals' human capital decisions and the emergence of teacher observation data. *Educational Researcher, 44*(2), 96–104.

Gomer-Mejia, L. R., Balkin, D. B., & Cardy, R. L. (2015). *Managing human resources* (8th ed.). Pearson Prentice Hall.

Gonzales, J. (2016, April 27) Tennessee terminates contract with TNReady test company. *Tennessean.* Retrieved from https://www.tennessean.com/story/news/education/2016/04/27/tennessee-terminates-contract-tnready-test-company/83594318/

Gonzales, J. (2018, August 7). Nashville and Shelby County schools superintendents declare "no confidence" in TNReady. *Commercial Appeal.* Retrieved from https://www.commercialappeal.com/story/news/education/2018/08/07/tnready-mnps-shelby-county-no-confidence/922718002/

Koedel, C., Li, J., Springer, M. G., & Tan, L. (2017). The impact of performance ratings on job satisfaction for public school teachers. *American Educational Research Journal, 54*(2), 241–278.

Koedel, C., Li, J., Springer, M. G., & Tan, L. (2019). Teacher performance ratings and professional improvement. *Journal of Research on Educational Effectiveness, 12*(1), 90–115.

Kupermintz, H. (2003). Teacher effects and teacher effectiveness: A validity investigation of the Tennessee value added assessment system. *Educational Evaluation and Policy Analysis, 25*(3), 287–298.

Loewus, L. (2011, October 18). Teacher-evaluation rush may jinx other state efforts. *Education Week.* Retrieved from https://www.edweek.org/policy-politics/teacher-evaluation-rush-may-jinx-other-states-efforts/2011/10

McQueen, C. (2015, March 1). How revised teacher evaluation measure has changed. *Tennessean.* Retrieved from https://www.tennessean.com/story/opinion/contributors/2015/03/01/tennessee-teacher-evaluation-system-explanation/24230435/

Olson, L. (2018). *Scaling reform: Inside Tennessee's statewide teacher transformation.* FutureEd, Georgetown University. Retrieved from https://www.future-ed.org/wp-content/uploads/2018/10/FutureEdTennReport.pdf

Papay, J. P., Taylor, E. S., Tyler, J. H., & Laski, M. E. (2020). Learning job skills from colleagues at work: Evidence from a field experiment using teacher performance data. *American Economic Journal: Economic Policy, 12*(1), 359–388.

Papay, J. P., & Laski, M. E. (2018). *Exploring teacher improvement in Tennessee: A brief on reimagining state support for professional learning.* Nashville, TN: Tennessee Education Research

Alliance. Retrieved from https://peabody.vanderbilt.edu/TERA/files/Exploring_Teacher_Improvement.pdf

Pepper, M. J., Burns, S. F., & Springer, M. G. (2012). *Educator evaluation in Tennessee: Preliminary findings from the 2012 first to the top survey.* Nashville, TN: Tennessee Consortium on Research, Evaluation, and Development. Retrieved from https://peabody.vanderbilt.edu/TERA/files/Educator_Evaluation_in_Tennessee_Preliminary_Findings_from_the_2012_First_to_the_Top_Survey.pdf

Rodriguez, L. A., Swain, W. A., & Springer, M. G. (2020). Sorting through performance evaluations: The influence of performance evaluation reform on teacher attrition and mobility. *American Educational Research Journal, 57*(6), 2339–2377.

Sanders, W. L., & Horn, S. P. (1998). Research findings from the Tennessee value-added assessment system (TVAAS) database: Implications for educational evaluation and research. *Journal of Personnel Evaluation in Education, 12*(3), 247–256.

Sanders, W. L., Saxton, A. M., Horn, S. P., & Millman, J. (1997). The Tennessee value-added assessment system: A quantitative outcomes-based approach to educational assessment. Grading teachers, grading schools: Is student achievement a valid evaluation measure? Thousand Oaks, CA: Corwin Press.

Springer, M. G., Swain, W. A., & Rodriguez, L. A. (2016). Effective teacher retention bonuses: Evidence from Tennessee. *Educational Evaluation and Policy Analysis, 38*(2), 199–221.

Swain, W. A., Rodriguez, L. A., & Springer, M. G. (2019). Selective retention bonuses for highly effective teachers in high poverty schools: Evidence from Tennessee. *Economics of Education Review, 68*, 148–160.

Tenn. Code Ann. § 49-5-501-515 (2012).

Tennessee Comptroller of the Treasury. (2018). *New audit examines TNReady testing failures* [Press Release]. Retrieved from https://comptroller.tn.gov/news/2018/12/19/new-audit-examines-tnready-testing-failures-.html

Tennessee Department of Education. (2013). *Partial year exemptions.* Retrieved from https://team-tn.org/wp-content/uploads/2013/10/Partial-Year-Exemptions_2017.pdf

Tennessee Department of Education. (2016). *Teacher and administrator evaluation in Tennessee: A report on year 4 implementation.* Retrieved from https://team-tn.org/wp-content/uploads/2013/08/TEAM-Year-4-Report1.pdf

Tennessee Department of Education. (2018). *Reflections over time: Tennessee educator survey 2018 results in context.* Retrieved from https://www.tn.gov/content/dam/tn/education/data/data_survey_report_2018.pdf

Tennessee Department of Education. (2019a). *2018–19 Educator evaluation composite weightings.* Retrieved from https://www.nctq.org/dmsView/Composite_weightings

Tennessee Department of Education. (2019b). *Lessons from our educators: Tennessee educator survey 2019b results in context.* Retrieved from https://www.tn.gov/content/dam/tn/education/data/2019b-survey/Survey_Report.pdf

Tennessee Department of Education. (2020a). *2020a–21 Achievement measure worksheet.* Retrieved from https://team-tn.org/wp-content/uploads/2020a/10/2020a-21_Achievement_Measure_Worksheet.pdf

Tennessee Department of Education. (2020b). *COVID-19 guidance: Impact on educator evaluation.* Retrieved from https://www.tn.gov/content/dam/tn/education/health-&-safety/Teacher%20and%20Administrator%20Evaluation_COVID-19_Guidance_4.21.20.final.pdf

Tennessee Department of Education. (n.d.-a). *General educator rubric.* Retrieved from https://team-tn.org/wp-content/uploads/2013/08/TEAM-General-Educator-Rubric-2018-19.pdf

Tennessee Department of Education. (n.d.-b). *Tennessee educator acceleration model (TEAM): Frequently asked questions: Level of overall effectiveness (LOE).* Retrieved from https://team-tn.org/wp-content/uploads/2020/08/TEAM-LOE-FAQ.pdf

Tennessee Department of Education. (n.d.-c). *Level of overall effectiveness (LOE) nullification guidance.* Retrieved from https://team-tn.org/wp-content/uploads/2013/10/final-loe-guidance.pdf

Tennessee General Assembly. (n.d.) *SB7005 actions/HB7010 actions* [Archived 106th General Assembly Bills and Resolutions]. Retrieved from http://wapp.capitol.tn.gov/apps/archives/default.aspx?year=106

Tennessee State Board of Education. (2020a). *Teacher and administrator evaluation policy (5.201).* Retrieved from https://www.tn.gov/sbe/rules-policies-and-guidance/policies.html

Tennessee State Board of Education. (2020b). *Educator licensure policy (5.502).* Retrieved from https://www.tn.gov/sbe/rules-policies-and-guidance/policies.html

Trout v. Knox County Board of Education, 163 F. Supp. 3d 492 (E.D. Tenn. 2016).

U.S. Department of Education. (2009a). *Race to the Top program: Executive summary.* Washington, D.C.: Author. Retrieved from https://www2.ed.gov/programs/racetothetop/executive-summary.pdf

U.S. Department of Education. (2009b, December 17). *States who have submitted letters of intent to apply for phase 1* [Press Release]. Retrieved from https://www2.ed.gov/programs/racetothetop/intent-to-apply.html

U.S. Department of Education. (2010a, March 29). *U.S. Department of education Arne Duncan's statement on race to the top phase 1 winners* [Press Release]. Retrieved from https://www.ed.gov/news/speeches/us-secretary-education-arne-duncans-statement-race-top-phase-1-winners

U.S. Department of Education. (2010b, March 29). *Delaware and Tennessee win first race to the top grants* [Press Release]. Retrieved from https://www.ed.gov/news/press-releases/delaware-and-tennessee-win-first-race-top-grants

U.S. Department of Education. (2013, August 9). *Statement by U.S. secretary of education Arne Duncan on Tennessee making changes to teacher licensure policy* [Press Release]. Retrieved from https://www.ed.gov/news/press-releases/statement-us-secretary-education-arne-duncan-tennessee-making-changes-teacher-licensure-policy

Wagner v. Haslam, 112 F. Supp. 3d 673 (M.D. Tenn. 2015).

Winerip, M. (2011, November 6). In Tennessee, following the rules for evaluation off a cliff. *The New York Times.* Retrieved from https://www.nytimes.com/2011/11/07/education/tennessees-rules-on-teacher-evaluations-bring-frustration.html

Part III
Teacher Evaluation Systems Around the World: Latin America

Chapter 8
Teacher Assessment in Chile

Yulan Sun

Abstract Since 2003, a national teacher assessment system has been applied in Chile. This chapter describes the origin and purposes of the teacher assessment, its instruments, consequences, and some results. It also reports on validation studies of the program and illustrates its evolution over time based on changes introduced in the instruments and, especially, the enactment of a Teaching Career Law in 2016. The chapter gives an overview of the teacher assessment 19 years after its introduction, recognizing successes, limitations, and pending challenges, not only for the assessment itself, but specially for the educational system and teacher policy. The most important of those challenges is to achieve the formative purpose of contributing substantively to the professional development and improvement of pedagogical practices.

8.1 Background and Characteristics of the Assessment System

8.1.1 Origin, Purposes and Consequences

The Professional Teacher Performance Assessment System, locally known as the "teacher assessment" (TA) was implemented for the first time in 2003, after a long negotiation involving the Ministry of Education (MINEDUC), the Teachers' Association, and the municipalities (responsible for the management of public schools). The process was not without complexities and resistance from an important part of the teaching staff (Ávalos & Assael, 2006; Bonifaz, 2011). However, in June 2003,

Currently affiliated with the Center for Teacher Professional Development (CDPD) of the Faculty of Education, Diego Portales University, between 2003 and 2019 Yulan Sun Figueroa led the MIDE UC team that leads the implementation of the Teacher Assessment.

Y. Sun (✉)
Universidad Diego Portales, Santiago, Chile
e-mail: yulan.sun@udp.cl

J. Manzi et al. (eds.), *Teacher Evaluation Around the World*,
Teacher Education, Learning Innovation and Accountability,
https://doi.org/10.1007/978-3-031-13639-9_8

the three parties signed a written agreement, which established central aspects of the assessment system including its purpose, instruments, and consequences (Colegio de Profesores, 2003). This agreement was endorsed a month later by 63.1% of the teachers who participated in a national survey, and the following year Law 19.961 was approved, which created the evaluation system.

As defined in the tripartite agreement, the law establishes that the TA is intended to evaluate the performance of around 85,000 classroom teachers who work in municipal or public schools[1] and represent approximately 44% of the national teaching staff. It was proposed as a *professional* performance assessment focused on teaching as the central task of classroom teachers, which would be based on explicit reference standards and would have a fundamentally formative nature, although the system also included from its origin summative purposes. The teachers receive feedback about the strengths and weaknesses of their performance. For those who obtain poor results on the assessment, free ongoing training with professional development plans[2] is offered. Furthermore, there are serious consequences including job loss for those who demonstrate sustained poor performance.

The TA describes four levels of performance: Outstanding, Competent, Basic, and Unsatisfactory. The level obtained on the assessment defines both the consequences and the periodicity of the assessment. Teachers who obtain the Competent (minimum expected) or Outstanding level are evaluated after four years; those with Basic results must be evaluated every two years, and those who are deemed to have an Unsatisfactory level must be evaluated the following year. These last two groups are required to participate in professional development plans, and teachers who obtain an Unsatisfactory result in two successive evaluations or fail to reach the Competent level in three must be dismissed. This is a consequence of very high impact because a central benefit of working in public schools is high job stability: the teachers who work in these settings are governed by a particular labor regulation, which guarantees their permanence except for when very serious and proven reasons are presented.

Regarding the consequences of the system, it should be noted that the TA came into being in a national and international context strongly inclined toward accountability in education and under the impact of national measurements of learning achievements, which showed large gaps compared to what was expected in the quality of teaching. Seen from a distance, it seems difficult that the TA system would have achieved sufficient political consensus and legitimacy in public opinion if some form of consequence for teachers with sustained poor performance was not established. On the other hand, the definition of the instruments and the participation of teachers in different and important roles within the process surely contributed to moderate the

[1] In Chile, there are four types of establishments that receive resources from the state according to their administrative regime and institutional framework. The two most important, which cover almost 90% of the country's enrollment (98% of which receive state subsidies), are municipal (administered by municipalities) schools and subsidized private schools, which are privately owned and administered. TA only includes classroom teachers of the former.

[2] These plans are designed and managed by local authorities with resources provided by the Ministry of Education according to the number of teachers with low results in each municipality.

apprehensions of the teachers and helped to make the implementation of the program feasible.

8.1.2 Responsible Entities and Teachers Evaluated

The global coordination of TA resides in the Ministry of Education through the Center for Pedagogical Training, Experimentation and Research (CPEIP). In each municipality, the law establishes the figure of a communal coordination, which corresponds to the local educational authority and executes different tasks associated with the process, such as the registration of those who must be evaluated and the delivery of reports on results. In addition, regulations require that to carry out the assessment MINEDUC must receive technical advice from a university, which is selected through a public tender. This role has been carried out from 2003 until now by the Pontificia Universidad Católica de Chile, through its Measurement Center MIDE UC.[3] In practice, this entity has been responsible—in coordination with the municipalities and with a variety of institutions and providers—for executing most of the complex process involved in implementing the TA each year.

As mentioned, although the TA is a nationwide program, it only applies to classroom teachers who work in municipal or public schools, who are governed by the "teacher's code." This is in contrast to peers who teach in "private subsidized" schools to whom the same labor code applies as to the rest of the country's workers.

The TA population includes all classroom teachers who work at different grade levels and in different subjects, but the implementation of the system was gradual: Each year different groups entered by grade level and subject along with a growing number of the country's 346 municipalities (see Table 8.1). This translated into a gradual growth in the scale of the process, which helped to manage the technical demands (in terms of measurement) and the logistics of the process.

8.1.3 Performance Standards: The Framework for Good Teaching

The TA is standards based; this means that the performance of each teacher is evaluated in relation to pre-established criteria and not in comparison with other teachers. These criteria are defined in the Framework for Good Teaching (Ministry of Education, 2008). This document was widely disseminated and reviewed within the school system, and its content is based mainly on previous work that formulated standards for initial teacher training in Chile and on the framework developed by Danielson in the USA (Taut & Sun, 2014).

[3] In 2022, some components of the program have been assigned to another university.

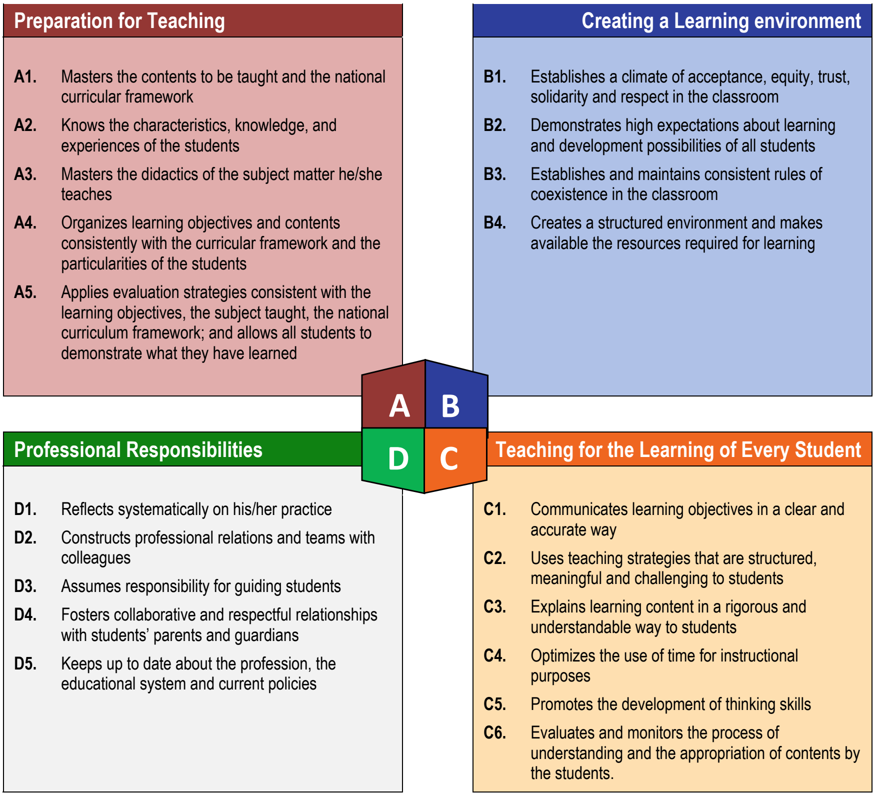
Table 8.1 Teachers evaluated by period according to level and teaching modality (*)

Year	Primary school (1° to 4º)	Middle school (5° to 8°)	Secondary school	Preschool (5–6 years of age)	Special education	Adult education	Technical/professional education	Total per year
2003	3673							3673
2004	920	799						1719
2005	8027	2291	347					10,665
2006	4937	8544	709					14,190
2007	3137	3650	3626					10,413
2008	2061	5556	4336	4061				16,014
2009	5164	4419	3630	900	1584			15,697
2010	3269	5193	1695	287	612			11,056
2011	3579	5293	1760	306	670	618		12,226
2012	2299	5165	4583	3021	985	362		16,415
2013	4421	5590	3486	1160	2092	307		17,056
2014	3639	6364	2790	1049	1960	237		16,039
2015	2949	5198	2880	676	1756	436		13,895
2016	3304	6309	4378	2328	2550	422		19,291
2017	4382	7018	4301	1497	4328	448	2267	24,241
2018	3729	7077	3549	1227	3704	377	592	20,255
2019	3266	5765	3109	956	3132	419	600	17,247

(*)Data only include teachers who complete the TA, not those from the subsidized schools who are only required to complete a portfolio as part of the teaching career recognition system

The TA instruments are based on the domains, criteria, and descriptors of the framework (see Fig. 8.1). This implies that the object of evaluation are fundamentally general pedagogical competencies common to the different levels and subjects, including only partial and indirect measures of knowledge of the discipline are included (e.g., in the review of the contents covered in the lessons or the learning assessments).

The Framework for Good Teaching was relevant for the acceptance of the TA because it guaranteed that the evaluation process would be based on explicit and valid criteria about the good performance of a classroom teacher. In this aspect, the Chilean system is aligned with the international recommendations as highlighted by a review of the program carried out by the OECD (Santiago et al., 2013). The review recommended that the framework be consolidated as a central pillar for teacher evaluation and professional development, be reviewed to correct some shortcomings, and updated in light of the most recent evidence from educational research. The results provided by the teacher assessment, which report strengths and challenges in teaching, should also serve as an input for the process. In 2014, MINEDUC



Preparation for Teaching

A1. Masters the contents to be taught and the national curricular framework

A2. Knows the characteristics, knowledge, and experiences of the students

A3. Masters the didactics of the subject matter he/she teaches

A4. Organizes learning objectives and contents consistently with the curricular framework and the particularities of the students

A5. Applies evaluation strategies consistent with the learning objectives, the subject taught, the national curriculum framework; and allows all students to demonstrate what they have learned

Creating a Learning environment

B1. Establishes a climate of acceptance, equity, trust, solidarity and respect in the classroom

B2. Demonstrates high expectations about learning and development possibilities of all students

B3. Establishes and maintains consistent rules of coexistence in the classroom

B4. Creates a structured environment and makes available the resources required for learning

Professional Responsibilities

D1. Reflects systematically on his/her practice

D2. Constructs professional relations and teams with colleagues

D3. Assumes responsibility for guiding students

D4. Fosters collaborative and respectful relationships with students' parents and guardians

D5. Keeps up to date about the profession, the educational system and current policies

Teaching for the Learning of Every Student

C1. Communicates learning objectives in a clear and accurate way

C2. Uses teaching strategies that are structured, meaningful and challenging to students

C3. Explains learning content in a rigorous and understandable way to students

C4. Optimizes the use of time for instructional purposes

C5. Promotes the development of thinking skills

C6. Evaluates and monitors the process of understanding and the appropriation of contents by the students.

Fig. 8.1 Domains and criteria of the framework for good teaching

began this review, still ongoing at the time of writing this chapter. However, in August 2019, the “Framework for Good Teaching in Early Childhood Education” was published (Subsecretaría de Educación Parvularia, 2019). Although the previous framework had been used as a reference for the assessment of preschool educators that work in the transition level (5–6-year-old children), there was a need to develop a specific framework that would take into consideration the particularities of this level of education and cover its entire cycle.

8.1.4 Final Results and Reports

The TA includes four instruments (detailed in the next section), and each has a weight, established by law, for the overall result: a self-assessment (10%), an interview conducted by a peer evaluator (20%), evaluation reports completed by school supervisors (10%, collectively), and a portfolio (60%). The same law modifies the weighting when a teacher has obtained an Unsatisfactory level in the previous evaluation: The portfolio increases to 80% and the other instruments lower their weight by half.

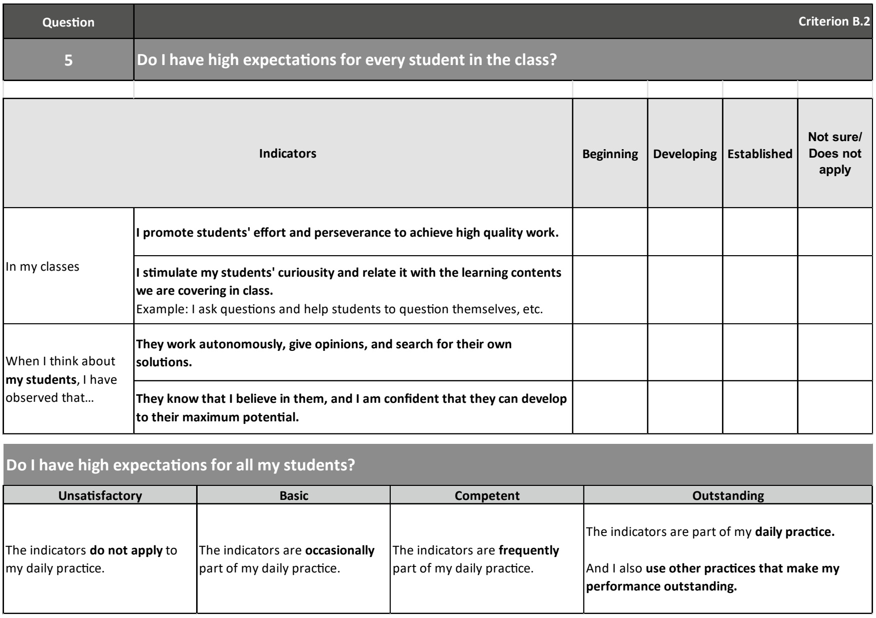
The weighted average of the instruments provides the global result. This is not necessarily the final result for each teacher because the law establishes that the final decision on this must be taken by a Communal Commission, under certain regulations (e.g., quorum of agreement). In fact, data have shown that these commissions, constituted in each community by peer evaluators and the local educational authority (the latter only with an opinion, but without the right to vote), confirm the results provided by the instruments in 95% of the cases.

As a result of the evaluation, each teacher receives an individual report that indicates his/her final and partial results, especially detailing performance on the indicators evaluated in the portfolio. In addition, the system provides reports with aggregated data for the school’s leadership teams and the local educational authorities, as an input for management, especially in terms of teacher professional development.

8.2 Assessment Instruments and Their Evolution

8.2.1 Self-Assessment

Teachers evaluate their own performance by answering a structured set of questions, common to all, on an online platform. In its current version, each question in the survey is made up of two parts. In the first part, the teacher is asked to analyze the degree of development reached in his/her teaching practice by four indicators that operationalize a specific criterion of the Framework for Good Teaching (FGT), while



Question	Criterion B.2
5	Do I have high expectations for every student in the class?

Indicators		Beginning	Developing	Established	Not sure/ Does not apply
In my classes	**I promote students' effort and perseverance to achieve high quality work.**				
	I stimulate my students' curiousity and relate it with the learning contents we are covering in class. Example: I ask questions and help students to question themselves, etc.				
When I think about **my students**, I have observed that...	**They work autonomously, give opinions, and search for their own solutions.**				
	They know that I believe in them, and I am confident that they can develop to their maximum potential.				

Do I have high expectations for all my students?

Unsatisfactory	Basic	Competent	Outstanding
The indicators **do not apply** to my daily practice.	The indicators are **occasionally** part of my daily practice.	The indicators are **frequently** part of my daily practice.	The indicators are part of my **daily practice.** And I also **use other practices that make my performance outstanding.**

Fig. 8.2 Question 5 of the 2020 self-assessment

in the second, each indicator must be scored with a performance level applying the rubric provided (see Fig. 8.2).

Self-assessment was included in the TA with the intention of promoting self-observation and analysis of one's own performance, but the results suggest that this purpose is not fulfilled, and the instrument has consistently shown a ceiling effect (see Fig. 8.9). This is not surprising given its weight in the final results of a high-stakes assessment.

Over the years, although the self-assessment has undergone variations, such as using more elaborate questions or asking for a rationale when self-assigning the Outstanding level, the changes have not shown any impact on the distribution of the results. It seems clear that the context of high consequences is decisive and has prevented the instrument from being the professional self-examination and reflection it was intended to be (Taut & Sun, 2014). At the same time, since the weight of the instruments is prescribed by law, any change (i.e., making the instrument an input for the analysis of the Communal Commission, but without weighting) would require a change in the legal framework.

8.2.2 Interview by a Peer Evaluator

A teacher of the same level and the same district (or one nearby) as the evaluated teacher conducts this interview using questions based on the Framework for Good Teaching. Approximately, 3000 teachers apply to be peer evaluators each year, and 40% are selected. These teachers complete a two-day training focused on how to apply the instrument and score responses based on rubrics. In general, the interview lasts around 50 min and takes place at the evaluated teacher's school. The evaluator must then enter the results in an online platform. As an example, in the 2020 interview one of the questions was "Can you describe a professional learning need that you have identified in relation to your pedagogical practice?" This question was linked to two criteria of the framework: D.1.2. critically analyzes teaching practice and reformulates it based on the learning outcomes of the students, and D.1.3. identifies learning needs and tries to meet them.

In their analysis of the TA, Santiago et al. (2013) recognized the value of including teachers themselves as evaluators and thus promoting development of competencies. Since 2003, to fulfill the role of evaluator, thousands of teachers have participated in trainings linked to the standards of the Framework for Good Teaching and the use of rubrics. They have also been able to learn more directly about the TA, which seems to have a positive impact on their perception of the program.[4]

On the other hand, the instrument has limitations, especially given that it is not based on direct evidence but on verbal statements from those evaluated, which may or may not reflect their actual practice. The OECD report also criticized the type of questions used and the fact that teachers did not know them in advance. Over time, improvements have been made to some of these aspects. For example, the type of questions was modified to refer to the *actual* practice of the teachers and not to their ability to recall the content of the Framework for Good Teaching. And, since 2014, the questions have been published in advance, which helps to reduce the anxiety associated with the interview and makes it easier for teachers to use them as a means to examine their own practices and reflect, individually or collectively, on them.

The change caused some concern among those responsible for the system regarding its effects on the scores because of the possible proliferation of responses prepared to achieve a high evaluation, but which did not reflect the real practice of the teacher. The changes did have an impact on the scores, but not exactly in that direction (see Fig. 8.3). The proportion of teachers at the Competent level increased, while that of the Outstanding level decreased. The result is interesting because it contradicts some conceptions within the educational system itself about evaluation, showing that knowing the questions in advance does not imply that those who lack competence will be able to demonstrate an ideal (fictitious) practice during the interview. Additionally, the publication of the questions was positively perceived by the teachers as demonstrated by their responses to the supplemental questionnaire. In it,

[4] For example, in a MIDE UC survey applied in 2012 to teachers who performed the role of peer evaluator in that year, 71% stated of them that the experience of being one had improved their opinion about the TA.

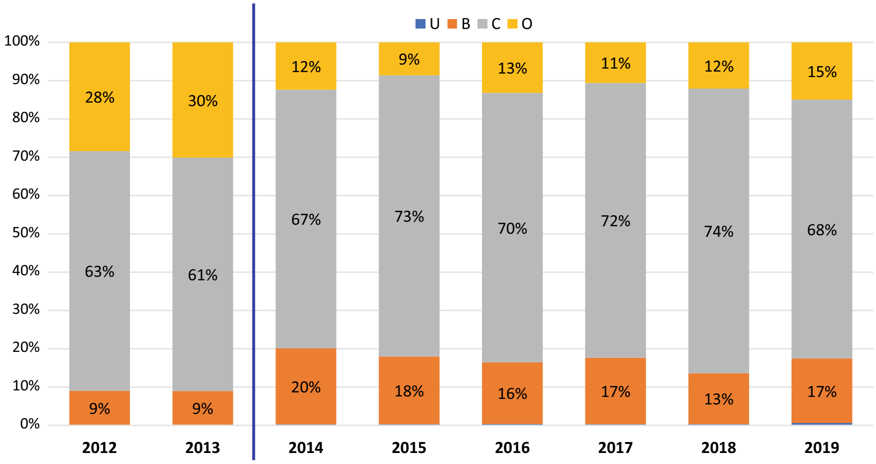


Fig. 8.3 Results of the interview by peer evaluator 2012–2019

92.4% stated that knowing the questions beforehand had been useful, for example, to reduce nervousness and anxiety before the interview, prepare their answers, and improve the quality of the conversation with their evaluator.

8.2.3 *School Supervisors Report*

This instrument is completed by the principal and the school's pedagogical coordinator, where the teacher evaluated works. As a way of promoting the use of self-assessment in the dialog and feedback between teachers and the school leadership, since 2012, the instruments that both complete have been made totally analogous in their content. In addition, seeking to strengthen the quality of their data, different changes have been tested in the instrument, for example, in the number and format of the questions and the scoring scale used. Among the changes, one included in 2010 stands out: When assigning the Outstanding level on a question, the school leadership was required to base the outstanding nature of that specific practice on a specific teacher. Without this justification, the question was scored as Competent. This change made a striking impact on the distribution of scores (see Fig. 8.4).

With this change, the average on the instrument dropped considerably: 0.3 points on a scale from 1 to 4 and a more restricted use of the Outstanding category was observed, concentrating the scores at the Competent level, a trend that has been maintained to date (Fig. 8.5). Thus, faced with the requirement to provide evidence of an Outstanding level performance, school leadership seems to be more reserved when evaluating teacher performance, using this level for those performances that were really considered to demonstrate excellence.

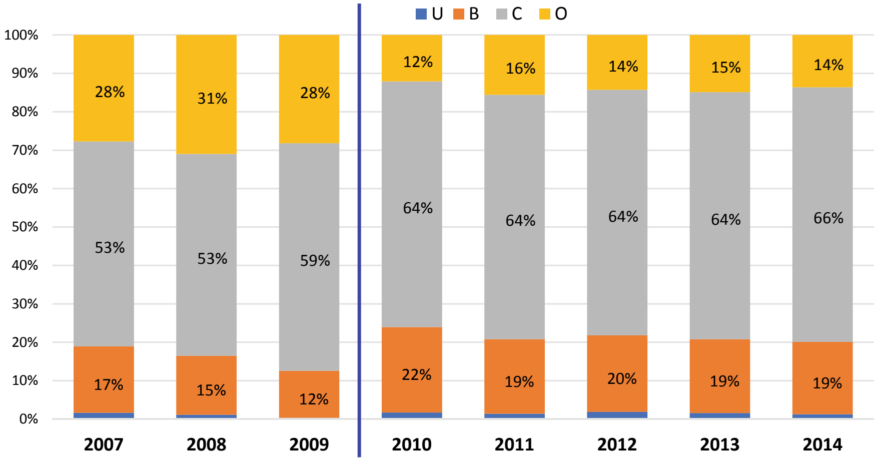


Fig. 8.4 Results of the school supervisors report before and after the change introduced in 2010

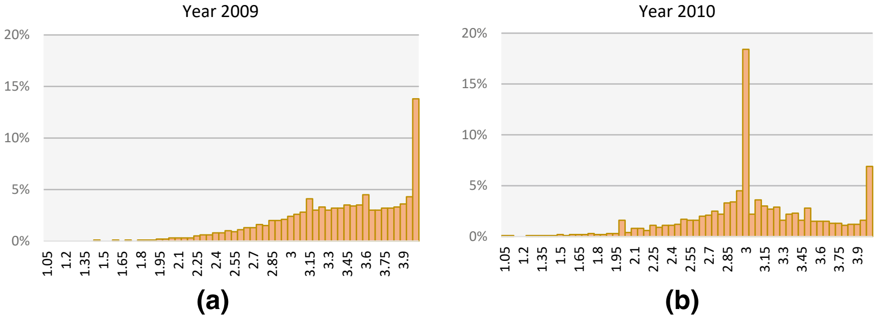


Fig. 8.5 Distribution of school supervisors scores before (**a**) and after (**b**) the change introduced in 2010

On the other hand, some criticisms and challenges persist regarding the instrument; one is the lack of a performance evaluation of school leaders, which would better balance accountability for the quality of teaching in the school. The low weight in terms of the final result has also been criticized, which accentuates the already limited powers that Chilean public schools' principals have to hire or fire teachers. Because of this, in 2011 a law was enacted that gave school principals the power to fire up to 5% of teachers among those who had obtained a Basic or Unsatisfactory result on the TA. In practice, however, the use of this power has been very limited, which could be explained by different factors: the scarce number of cases that fall within that 5% and the interpersonal or "emotional" burden of firing a team member or the economic costs associated to it (Concha, 2015; Metropolitan Technical University, 2017).

One change indirectly linked to the instrument has been the incorporation of short training seminars for school leadership. Since 2010, approximately 800 principals and school pedagogic coordinators each year have participated in these experiences that have addressed topics such as strategies to complete the instrument, analysis and use of the Result Reports provided by the TA, classroom observation and feedback on teaching practices, and the knowledge of teaching quality indicators evaluated in the portfolio. Beyond their specific focus each year, these trainings allow school leaders to better understand the TA, and for those who implement the program, the seminars provide a direct channel of interaction with crucial actors for the improvement of teaching. These opportunities seem to play an important role in terms of mediating the impact of the TA on the teachers being evaluated (Sun et al., 2017) and the way in which the process is used constructively by the school (Taut et al., 2011a).

8.2.4 Portfolio

This is the instrument with the greatest weight for the overall result of the TA, and not only for that reason, it is a central element of the evaluation system. Using a portfolio to evaluate teacher performance on a national scale and with such high consequences was quite a pioneering experience within the region and a prominent technical challenge from the point of view of its construction and scoring. In Chile, its only antecedent was the portfolio used in the Assignment for Pedagogical Excellence program, which had started only one year before and which, because it was voluntary, had a different character and a very small scale.

Unlike the other instruments, which are based on reporting or self-reporting, the portfolio more directly measures performance and is closely related to what teachers do regularly in their work. Also, within the TA instruments, the portfolio involves the most sophisticated and controlled construction and correction processes. The tasks that define the evidence to collect as well as the rubrics to evaluate it involve an extensive process with the participation of assessment experts and teachers specialized in the different subjects and levels evaluated. In addition to the Framework for Good Teaching (FGT) as a theoretical base, the construction of the portfolio and its rubrics consider periodic updating of empirical evidence and relevant theoretical developments, as well as contextual factors (e.g., new regulations and educational policy directions) through bibliographic review and interviews with key informants; analysis of data from previous applications; qualitative studies in which instructions and tasks are tested and adjusted, along with pilot studies of evidence put together by the classroom teachers; analysis of the consultations and opinions expressed by those evaluated themselves through questionnaires and records from a call center and a web-based consultation service offered by the program. Information is also collected from the portfolio raters, which allows to identify areas of improvement in the rubrics, procedures, and scoring materials. These and other processes obey a rational of continuous improvement and search for validity in the elaboration of the instrument (Torres & Zapata, 2019).

Teachers have 12 weeks to complete their portfolio using an online platform. To do this, they receive a manual containing detailed instructions. The manual presents specifications according to the subject, level, and teaching modality based on a common structure.

The portfolio is organized into three modules (see Table 8.2). Module 1 collects evidence related to teaching planning, students learning assessment and teacher reflection on the learning process, and his/her practices. Module 2 consists of a videotaped lesson of 40 min, which allows teaching practices to be directly observed, including aspects such as classroom management, learning environment, the participation of students, and the way in which the teacher interacts pedagogically with them. Module 3 was incorporated in 2016 as a result of the by Teaching Career Law (see next section). In it, each teacher must give an account of a collaborative work experience in which he or she has participated to address problems or needs relevant to students learning. The presentation of Module 3 has been voluntary up until today, and for those who deliver it, their score is only considered if it benefits the overall result of the portfolio; otherwise, feedback is provided but has no effect on the portfolio score. This decision is due to the interest in promoting collaborative work in schools, while recognizing that in many cases there is no school culture and/or conditions for this type of interaction among teachers.

Portfolio correction is a fairly sophisticated process that involves multiple steps and stakeholders. Every year approximately 600 teachers act as portfolio raters, each one assessing portfolios of the level and subject area that correspond to their training and teaching experience. These teachers are prepared through a preliminary online course followed by a face-to-face 30-h training focused on the application of scoring rubrics on real evidence. A trial period is also implemented, which allows the appropriate functioning of the entire scoring process to be evaluated (without effect on the evaluated teachers' scores).

The task is carried out in scoring centers, housed in different universities but is centrally monitored through supervision in the field and online, including both the procedures developed and the data obtained. To promote the reliability of the process, in addition to training, mechanisms such as (blind) double scoring of 30% of the evidence[5] and "master coding" are implemented, in which each group of raters (by level/subject) qualifies the same evidence, and then a score analysis and discussion are conducted by the supervisor team.

Although research has shown that the portfolio is the instrument that provides the most information within the TA (see section on research and validation), it has not been exempt from criticism. One relevant criticism is that scoring rubrics have not been known by teachers or the public in general. Only in 2021, this changed, and rubrics were published in the program website.[6] Yet, although rubrics were not public as such in previous years, the Portfolio Manual described the Competent performance in all the indicators evaluated in this instrument, and this was based

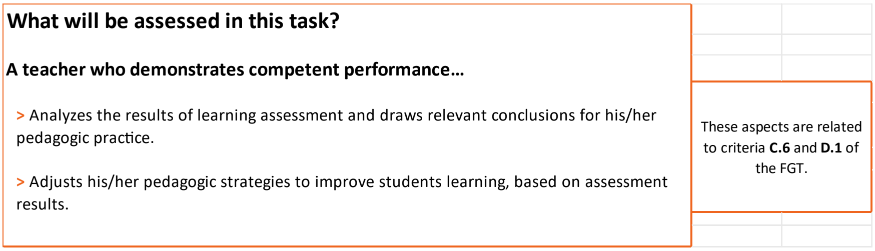
[5] Since the correction is done by module, this implies that for close to 50% of the teachers, at least one of the modules in their portfolio has been doubly corrected.

[6] https://www.docentemas.cl/portafolio/rubricas/.

Table 8.2 Modules and tasks of the 2020 portfolio

	Task	Subtask	Indicators evaluated
Module 1	Planning	– Description of three lessons of a learning unit	– Formulation of learning objectives – Relationship between learning objectives and activities
	Assessment	– Classroom assessment instrument – Analysis of results	– Learning assessment and scoring guidelines – Relationship between assessment and learning objectives – Analysis and use of learning assessment results
	Reflection	– Analysis of students' characteristics – Learning from error	– Analysis based on students' characteristics – Using error for learning
Module 2	Videotaped lesson	– Video of a 40-min lesson – Information about the lesson	– Classroom learning environment – Promoting students' participation – Quality of lesson opening – Quality of lesson closure – Contribution of classroom activities to learning goals – Implementing curriculum-specific directions according to grade level and subject – Quality of teacher's explanations – Quality of questions and activities – Feedback to students
Module 3	Collaborative work	– Description of a collaborative work experience – Reflection from the collaborative experience	– Relevance of the need or problem addressed by the collaborative action – Professional dialog in the collaborative action – Value of collaborative work for professional development – Reflection on the impact of the collaborative work experience

Note Prepared by the author based on the tasks and subtasks of the 2020 Portfolio Manual for middle school (fifth–eighth grade) teachers and the indicators reported in the Result Reports for the same year



What will be assessed in this task?

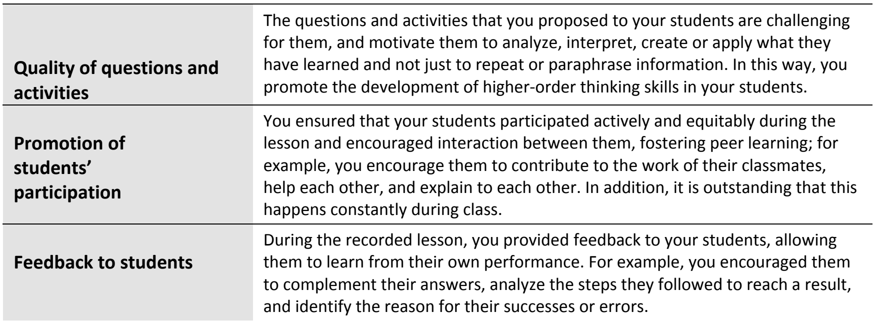
A teacher who demonstrates competent performance…

> Analyzes the results of learning assessment and draws relevant conclusions for his/her pedagogic practice.

> Adjusts his/her pedagogic strategies to improve students learning, based on assessment results.

These aspects are related to criteria **C.6** and **D.1** of the FGT.

Fig. 8.6 Example of the description of the Competent level in the 2020 middle school portfolio manual



Quality of questions and activities	The questions and activities that you proposed to your students are challenging for them, and motivate them to analyze, interpret, create or apply what they have learned and not just to repeat or paraphrase information. In this way, you promote the development of higher-order thinking skills in your students.
Promotion of students' participation	You ensured that your students participated actively and equitably during the lesson and encouraged interaction between them, fostering peer learning; for example, you encourage them to contribute to the work of their classmates, help each other, and explain to each other. In addition, it is outstanding that this happens constantly during class.
Feedback to students	During the recorded lesson, you provided feedback to your students, allowing them to learn from their own performance. For example, you encouraged them to complement their answers, analyze the steps they followed to reach a result, and identify the reason for their successes or errors.

Fig. 8.7 Example of feedback for indicators evaluated in the Module 2 of the portfolio (videotaped lesson)

directly on the rubrics (Fig. 8.6). Also, the feedback texts that the teacher receives in their Result Reports (Fig. 8.7) are directly extracted from the rubrics. In addition, on the TA website examples of the practices evaluated are presented, accompanied by a brief explanatory analysis. In this way, although publishing the rubrics should be considered a positive change, their content was not hidden from those who were evaluated.

8.3 Some Results: What the Evaluation Says About Teacher Performance

Although the cohorts evaluated each year have different compositions as a result of the gradual incorporation into the program and the periodicity rules according to the previous result, in rough terms the results profile shows some similarities. Most of those evaluated are concentrated in the level of "Competent," which corresponds to the minimum expected level, followed by the Basic level, while a small percentage reach the Outstanding level, and an even lower group is rated Unsatisfactory (Fig. 8.8). Therefore, the group of teachers forced to leave their job because of

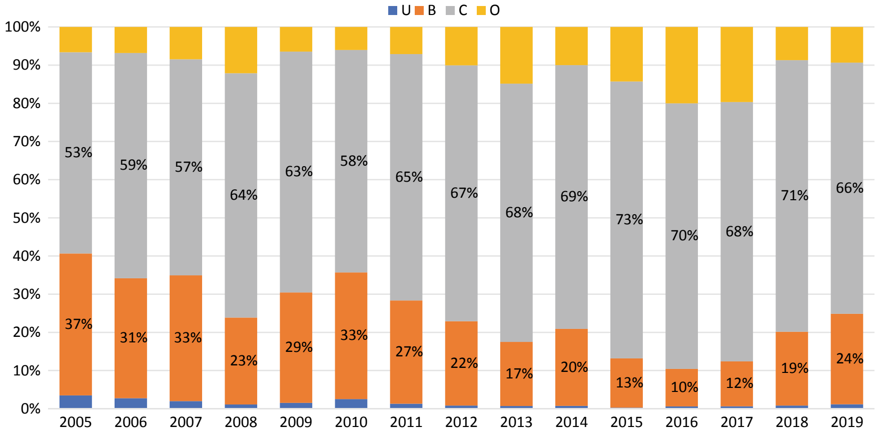


Fig. 8.8 Distribution of final result in the TA 2005–2019[7]

their results on the TA is a minimal proportion of the total evaluated: 984 teachers in 17 years, from the start of the process until 2019.

However, the profile of final results hides substantive differences between the instruments: These show different distributions of teacher performance, which have also been quite stable over time (Fig. 8.9). A variety of factors could explain this heterogeneity. High consequences certainly play a role especially in self-assessment, but also among the supervisors' report and peer evaluators. In training, the latter often express their reluctance to assign low grades, which may harm their colleagues in a way that they consider unfair. The portfolio, on the other hand, has a highly controlled, anonymous scoring system, and it is based on an assessment of more direct evidence of teaching work.

Due to its characteristics, the portfolio seems to provide a more reliable and informative description of the strengths and weaknesses of teaching performance (Fig. 8.10). Among the former are practices related to planning, such as the formulation of learning objectives, the coherence between those objectives and the activities designed to meet them, and the ability to promote students' participation in which most teachers reach or exceed the Competent level (91%, 65%, and 60%, respectively). On the other hand, the lowest results are obtained on indicators related to the use of errors for learning, the learning environment in the classroom, and different aspects of pedagogical interaction, such as the quality of the questions and activities proposed to the students and the feedback provided to them by the teacher. On these indicators, a reduced proportion of teachers (12% to 20% depending on the case) reach or exceed the Competent level.

[7] Data from 2020 and 2021 have not been included due to the anomalies in the process as a result of the health contingency associated with the COVID-19 pandemic. For this reason, not only was the number of those evaluated reduced very substantially, but also the process had to undergo multiple adjustments with respect to previous periods.

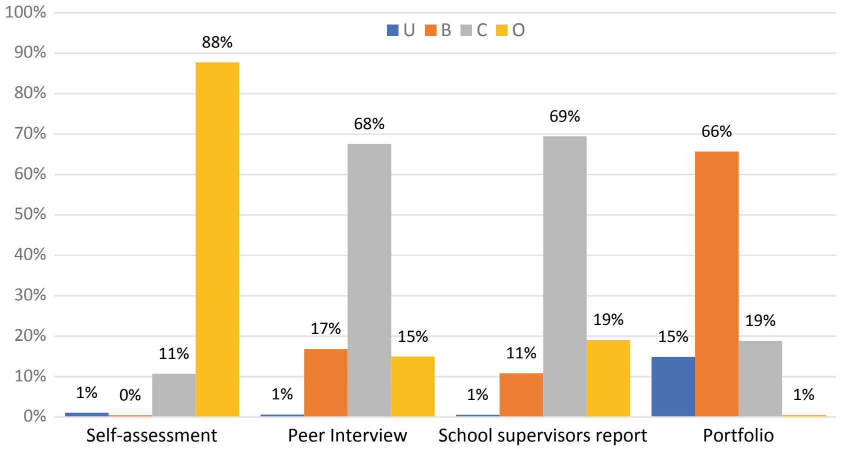


Fig. 8.9 Results by TA instrument 2019

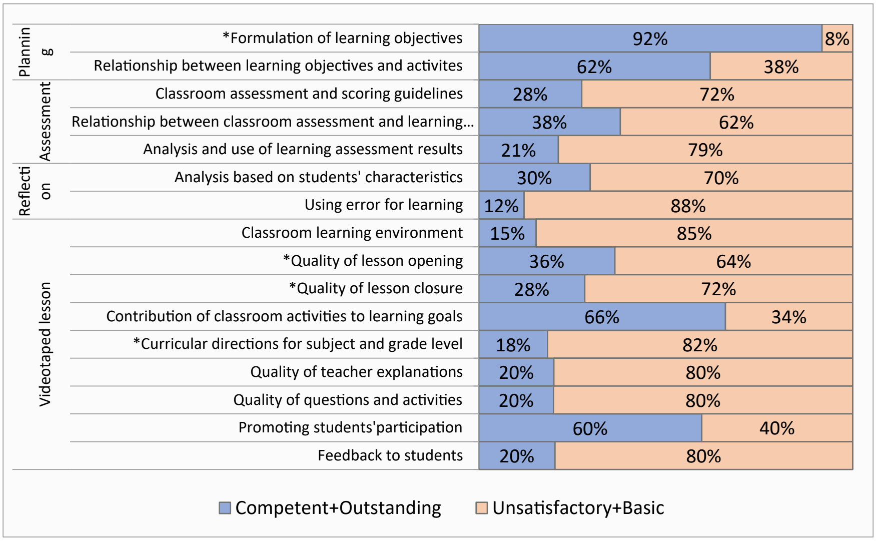


Fig. 8.10 Result by indicator evaluated in Modules 1 and 2 of the 2019 portfolio. *Notes* Indicators only used to evaluate teachers of specialty areas in Technical Professional Secondary Education are omitted. The asterisks mark indicators that are not evaluated for these teachers but are evaluated for the rest

In the case of Module 3, which the majority of teachers voluntarily present (79% in 2018 and 66% in 2019), the results are generally low (Fig. 8.11). Only a third of those evaluated report an experience clearly aimed at improving the learning of their students, and an even smaller proportion manages to reflectively analyze the impact of that experience on the educational community and on their own practice.

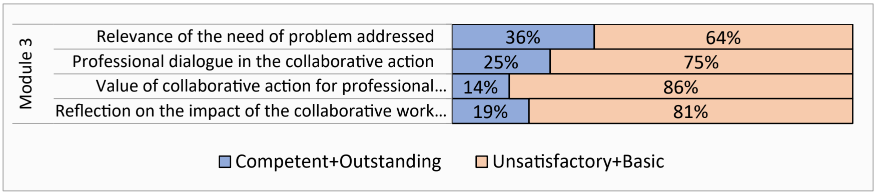


Fig. 8.11 Result by indicator evaluated in Module 3 of the 2019 portfolio

The results corroborate that collaborative work is still an incipient practice in the Chilean educational system. Including it in the Portfolio can help to underscore its importance, communicate its meaning and characteristics, and promote conditions so that teachers can undertake collaborative work in their schools. At the same time, it is a new challenge for an instrument and a system that until now have focused on individual practice.

8.4 Research and Validation

The participation of a university in the implementation of the TA probably facilitated the generation of a fairly comprehensive validation agenda around the program (see, for example, Taut et al., 2011b; Taut et al, 2012a).[8] Furthermore, a variety of studies and publications have analyzed it from different angles, i.e., its origin and installation (Avalos & Assael, 2006), its contribution to teacher professional development (Avalos-Bevan, 2018; Roa-Tampe, 2017), and the significance, perceptions, and representations of different actors regarding the evaluation (Fardella & Sisto, 2015; Roa-Tampe, 2018; Rosales, 2018; Sepúlveda et al., 2019; Sisto et al., 2013; Sun et al., 2017; Tornero & Taut, 2010; Urriola, 2013). Also, in 2011, the Ministry of Education commissioned a panel of experts convened by the OECD, which analyzed the TA in detail, its governance, instruments, implementation, and effects, identifying its strengths and pending challenges (Santiago et al., 2013). The analysis of these studies far exceeds the scope of this chapter, which will focus on describing some of the most interesting findings regarding the validation of the system.

8.4.1 Classification of Teacher Performance

Two studies have investigated the consistency between the performance classification given by the TA and other measures of teacher quality, such as teacher knowledge

[8] Although the researchers who have led this agenda belong to the institution that advises the teacher evaluation, they are not part of the implementation team, and the studies have been carried out within the framework of competitive funding, following the standards required by academic research.

measured through a written test, classroom practices evaluated through direct observations or through videos (different from those presented for the TA), analysis of pedagogical materials, questionnaires to students about their perception of teaching practices, and measurement of student achievement at the beginning and end of the school year. The first study focused on the extreme categories, looking at whether the new measures confirmed the differences indicated by the TA between teachers with Outstanding and Unsatisfactory final results (Santelices & Taut, 2011). The second study—carried out several years later—undertook a similar analysis, but this time with the intermediate categories, Competent and Basic (Taut et al., 2019). In both cases, differences were found consistent with the performance classification made by the TA.

8.4.2 Relationship Between Results on the TA and Student Learning Achievements

This relationship has been explored using different methodologies. From a descriptive point of view, several SIMCE[9] reports have showed that students who have had a greater number of high-performing teachers (Competent or Outstanding) obtain higher achievements on this measure (Taut & Sun, 2014). Other studies have linked the result on the TA with the achievement of the students in a more direct way, but only cross-sectional (crossing data from a specific point in time) or using aggregated data from both measurements. Their results also provide support for the relationship between the achievement of students in SIMCE and the result of teachers on the TA, especially with the Portfolio (see, for example, Alvarado et al., 2012; Bravo et al., 2008; Eisenberg, 2008; León, 2008; Manzi et al., 2008).

Using hierarchical linear models (HLM), Santelices and Taut (2011) analyzed longitudinal data on the learning of students who had teachers with Outstanding and Unsatisfactory results. They found that the classification on the TA is a significant predictor of student achievement at the end of the school year, controlling for initial baseline data. Later, the analysis of the relationship between the result on the TA and the progress in students' performance showed that indices of added value for mathematics teachers and, to a lesser extent, for language teachers were significantly correlated with their performance on the TA, and especially with those of the portfolio (Taut et al., 2012b, 2014). In summary, studies have consistently indicated that there is a relationship between performance on the TA, especially on the portfolio, and student learning outcomes.

[9] Quality of Education Measurement System (national standardized tests for measuring learning achievements).

8.4.3 Evidence of Consequential Validity

Different studies have investigated to what extent the intended effects of the evaluation system are fulfilled, including the professional development plans associated with the results (Cortés et al., 2011), the participation of teachers in a program of incentives linked to the process (the VAIP),[10] and the teachers' career paths, including the probability of leaving their jobs according to their results on the TA (Taut et al., 2010). The perceptions of relevant actors, including teachers, the school leadership team, local authorities, and MINEDUC officials, about the intended and unintended consequences of the program were also researched (Taut et al., 2011a; Santelices et al., 2013). The data reveal heterogeneity in the degree to which the expected consequences are met (see Table 8.3).

Research also shows unintended effects, both positive and negative. Among the former are the support provided by schools and municipalities to teachers on their evaluation period and the training impact that the TA triggers in different ways, for example, by promoting knowledge and analysis of the Framework for Good Teaching or through the experience of developing the portfolio. Among the negative effects are the work overload that teachers experience when being evaluated, negative emotional reactions (such as anxiety and job insecurity) that accompany the process, and the emergence of fraudulent or unethical practices, such as evading the obligation to be evaluated using legal subterfuge or copying and buying portfolios.

Table 8.3 Summary assessment of empirical findings regarding the TA's intended effects (adapted from Taut & Sun, 2014)

Intended uses	Evidence
Ranking teachers according to their performance	+
Diagnose strengths and weaknesses in teachers' practices	+
Strengthen collaboration between teachers	+
Provide information for decision-making at the local level	(+)
Promote the social recognition of high-performing teachers	(+)
Improve job prospects through monetary incentives (VAIP)	0
Support professional development through professional improvement plans	0

Note + *indicates substantial or consistent evidence; (+) limited or heterogeneous evidence; 0 no evidence was found*

[10] The Variable Assignment for Individual Performance (VAIP) is an economic incentive that could be obtained by teachers who, having performed well on the TA, also took a knowledge test and obtained good results. Following the enactment of the Teaching Career Act, this payment ceased to exist.

8.4.4 *Experiences, Representations, and Teaching Discourses Around the TA*

Several studies complement from a more hermeneutic perspective, the results presented above, shedding light on the way in which school actors, especially teachers, experience the TA, and the representations and rhetoric they build around it. This is relevant because of the undoubted influence that these processes have on the impact of educational policies. For this reason, and for illustrative purposes (which are in no way exhaustive), some their results will be briefly mentioned.

Roa-Tampe (2018) analyzed the rhetoric of 40 teachers regarding the TA and the Framework for Good Teaching. Among her findings, she points out that an important part of the rhetoric constructs a perception of illegitimacy of the TA, understood as an exogenous regulation and dismissing it as a "fiction," far from the practice and real contexts of teachers. This author also notes that, to face their evaluation, teachers develop peer support practices, thus building a group rationality to face a process that is defined as individual. The scope of this collegiality, in any case, seems to be limited as it approaches the TA on the basis of linguistic techniques and keys (terms or formulas that would lead to a good result), rather than taking a thoughtful, genuine, and professional approach.

Fardella and Sisto (2015) analyzed the discourse of interviews to 20 teachers and found that when confronting the categories posed by the TA, they develop processes of *subjective ascription* as well as *subjective resistance*. Both of these processes show that teachers' rhetoric does not reproduce linearly and unequivocally the official discourse of the policies of strengthening the profession, but rather builds their own in a local and heterogeneous process. This is expressed, for example, in the fact that teachers welcome and apply some of the categories and distinctions of the TA (in part because of the need to give it intelligibility); but at the same time, they develop practices of questioning and justification, such as disputing these categories, declaring their discomfort or demanding the recognition of other aspects of themselves as teachers. In this way, teachers seek to preserve a subjectivity stressed by exogenous distinctions and to diminish the logic of control and surveillance that would underlie the evaluation system.

Another study based on 42 interviews with teachers and principals from 13 schools confirms the tensions that arise around the TA: within the program itself, for its dual summative and formative purpose, and also in teachers experience (Sun et al., 2017). Concerning the program, the encouragement for teachers to analyze and show their practice *as it is* (in a genuine way) brings with it the possibility of being harmed by the consequences of a bad result. Then, several "adjustments" are observed, which imply distancing oneself from the authentic practice, hence compromising the possibility of reflection and feedback. As for the teachers' experience, they reveal a professional identity strongly based on the practical, rooted in their particular context and focused on responding to the social and emotional needs of their students. In contrast, they perceive the TA as a system from the realm of "the theoretical," based on common standards (insensitive to context) and omitting relevant areas of their work. The study also shows that some teachers do see the evaluation as an experience of reflection and professional learning, but the small number of them confirms both the potential of the process and its internal difficulties to fulfill a developmental goal.

8.5 The Introduction of the Teacher Professional Development Law (2016)

Establishing a *teaching profession* was an aspiration already present in the origins of the TA, but ultimately this did not materialize in the agreement that gave rise to the program (Avalos-Bevan, 2018). Over time, several legal changes have complemented or modified the standards that regulate the TA, but without returning to the idea of a teaching profession. Thus, for example, as a way of incorporating disciplinary and pedagogical knowledge (which the TA does not measure directly), in 2004 an economic incentive was created for teachers who, having had a good result on the TA, also obtained good results on a test of disciplinary and pedagogical knowledge (Law 19.933, 2004). In 2005, the situation of those teachers who should have been evaluated but refused to do it was regulated, establishing that they would be presumed to have an Unsatisfactory result. And in 2011, a rule was enacted (Law 20.501) that hardened the consequences of the TA, defining that a teacher should be fired not after three consecutive results with an Unsatisfactory level, but after two, and the same if in three consecutive evaluations he/she did not reach the Competent level.

Finally, in April 2016, 16 years after the introduction of the TA, Law 20.903 was enacted, creating the National Teacher Professional Development System. This law, also known as the "teacher career law," represents the most important change in Chilean teaching policy in the last 25 years and addresses multiple areas regarding the training and exercise of the teaching profession; it includes new requirements for students entering it and for training programs, mentoring for novice teachers, changes in the vision and management of teacher professional development, and reduction of teaching hours for classroom teachers, among others. In addition, the law creates a recognition system that establishes professional development levels, to which a new scale of wages is associated. According to estimates of the MINEDUC, the introduction of this system implies, on average, an increase of 30% in a teacher monthly salary, with a possibility to even double it.[11] With this, Chile addressed a long-delayed need to improve teacher salaries, whose disparity with those of other professions was an inconsistency increasingly difficult to accept and a widely recognized obstacle to improving the attraction to teaching profession and the quality of its practice.

The recognition system identifies five stages: Initial, Early, Advanced, Expert I, and Expert II. Each teacher is allocated in one of them based on their professional

[11] It should be noted that the same law ensured that no teacher would see his/her salary reduced due to the evaluation process and its results, but it may not increase either.

Table 8.4 Stages of the recognition system according to results on the instruments

	Knowledge test			
Portfolio	A	B	C	D
A	Expert II	Expert II	Expert I	Early
B	Expert II	Expert I	Advanced	Early
C	Expert I	Advanced	Early	Initial
D	Early	Early	Initial	Initial
E	Initial			

experience (years of practice),[12] subject and pedagogical knowledge (assessed by means of a test), and pedagogical competencies (evaluated through a portfolio, the one already used in the TA, adding the module on collaborative work). The stage thus defined (see Table 8.4) is the most determining factor in the salary of each teacher and, to remain practicing, she/he must reach the Advanced level. The two higher stages, on the other hand, are optional (voluntary), but financial incentives have been put in place to encourage teachers to reach them.

Along with the dramatic change in the consequences, another relevant implication of Law 20.903 is the expansion of coverage of the TA throughout the teacher work force; since the recognition system considers all teachers who work in schools that receive resources from the state, it will eventually cover 90% of the national teaching staff.[13] Also it will incorporate about 51,000 professionals responsible for the education of infants and children up to the age of 4, whose technical and administrative dependence is different from that of schools. In this way, the teaching career will reach almost universal coverage, leaving out only private education. Therefore, the impact of the new evaluation process is undoubtedly a theme to be analyzed and studied. The introduction of such direct and substantive consequences on teachers' salaries is unprecedented in Chile, and research should be done to monitor the development of the system and to what extent the desired objectives are achieved.

8.6 Final Remarks

Overall, the experience of the Chilean TA system could rightly be described as "successful" or exemplary in several respects. First, there is the complex, but fruitful negotiation between three actors who usually take opposing positions. Also, the choice of a standards-based assessment model helps to make explicit and socialize what is meant by quality teaching. The decision to give a leading role to an instrument

[12] For example, the Advanced level can only be reached after 4 years of experience and that of Expert I with 8.

[13] The TA, on the other hand, only includes municipal schools, which account for 34.3% of school enrollment and 43.5% of the country's teaching staff (Statistical Yearbook 2018 published by the Ministry of Education in 2019).

such as the portfolio, better equipped to address the complexity and contextual nature of teaching, was also a plus. And the development of a robust body of research has provided evidence on the validity of the program and its instruments. The continuous improvement of the program, which has combined the stability given by the law with a sensitivity to feedback and advances in knowledge, is another positive aspect. Finally, it should not be ignored its trajectory of almost two decades, under administrations of different political orientations and in sometimes very complex contexts, including natural disasters such as the earthquake that hit Chile in 2010.

At the same time, and perhaps more clearly because of its virtues, the experience also reveals the limits of a system like this to enhance professional development and thereby to improve teaching. With all its achievements, an unfulfilled promise persists in the TA in this regard (Avalos-Bevan, 2018; Roa-Tampe, 2017; Santiago et al., 2013; Sun et al., 2017). The quality of teaching and learning has not changed substantially and remains well below expectations. After 19 years, it is reasonable to ask how much more can be expected, for this purpose, from the TA or the new Recognition System: Is it feasible to promote reflection and professional learning in the context of an evaluation of such high consequences? Can these systems drive the great change in education that the country needs and longs for?

The difficult coexistence between summative and formative purposes seems to be resolved, in fact, with the primacy of the former. In my opinion, this problem is at a level that goes beyond the evaluation system itself, and therefore, it will not find a solution in any changes or improvements to its instruments, performance reports, or other devices. Without denying the value that these changes may have, they all occur within the same logic, which in turn impacts the perception that teachers and other actors have of the evaluation process and how they face it. It has already been described how, even in a stage of undeniable consolidation of the program, in the representations and rhetoric of the school actors about the TA, there is a level of distrust, questioning and conflict with teachers' professional identity. This in turn results in an often bureaucratic and/or "strategic" approach to the assessment process, which inevitably diminishes its formative potential.

The new consequences probably accentuate this context, which does not encourage to venturing into a process of self-observation, reflection, and learning. This does not mean that the TA is unnecessary, but that it is insufficient to fulfill certain purposes. And, although more recent, something similar could be expected from the Recognition System, since its consequences, although positive (salary increase), are even stronger and its impact more extensive. In contrast to this system, TA has no direct association with economic incentives and its most drastic consequences have affected an extremely small group of teachers, because they are linked to the overall result, not exclusively to the portfolio.

Law 20.903 offers an opportunity to advance over and beyond the TA and the Recognition System. It presents a vision that better reflects the value and complexity of the teaching profession, the collegial nature of its exercise, and the importance of reflective practice and collaboration for lifelong professional learning, among others. So far it is clear that the most visible impact of the law—at least at the level of practicing teachers—has been in the Recognition System. This has powerfully focused the attention, leaving formative evaluation and professional development dangerously in the background and does not bode well for the lessons learned after nearly two decades of teacher evaluation.

In my opinion, the quality and equity of education in our country require taking new and qualitatively different steps: a combination of *trust, support,* and *rigor* and a perspective that is more and better founded on pedagogy; after all, it is a case of *learning*, in this case on the part of the teachers. As far as evaluation is concerned, these steps require a different perspective, based on a pedagogical understanding of the issues and problems at hand, decidedly formative in its purpose, concentrating resources and efforts on *support* rather than on *measurement*.

References

Alvarado, M., Cabezas, G., Falck, D., & Ortega, M. E. (2012). La Evaluación Docente y sus instrumentos: discriminación del desempeño docente y asociación con los resultados y asociación con los resultados de los estudiantes. MINEDUC-Programa de las Naciones Unidas para el Desarrollo. Santiago, Chile. http://www.cl.undp.org/content/dam/chile/docs/pobreza/undp_cl_pobreza_informe_doc6.pdf

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

Avalos-Bevan, B. (2018). Teacher evaluation in Chile: Highlights and complexities in 13 years of experience. *Teachers and Teaching, 24*(3), 297–311. https://doi.org/10.1080/13540602.2017.1388228

Bonifaz, R. (2011). Origen de la Evaluación Docente y su conexión con las políticas públicas en educación. In J. Manzi, R. González, & Y. Sun (Eds.). *La Evaluación Docente en Chile.* Santiago, Chile: MIDE UC. https://www.mideuc.cl/libroed/pdf/La_Evaluacion_Docente_en_Chile.pdf

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

Concha, M. C. (2015). Criterios complementarios a la evaluación de desempeño profesional docente que prevalecen en directores para desvincular a docentes del sistema municipal: estudio descriptivo en 3 escuelas básicas vulnerables. Tesis para optar al grado de Magister en Ciencias de la Educación, Pontificia Universidad Católica de Chile. https://repositorio.uc.cl/handle/11534/15782

Colegio de Profesores de Chile A.G. (2003). Fin de las calificaciones: Nuevo Sistema Nacional sobre Evaluación del Desempeño Docente. Revista Docencia No 20 (pp. 80–87).

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

Eisenberg, N. (2008). *The performance of teachers in Chilean public elementary schools: exploring its relationship with teacher backgrounds and student achievement, and its distribution across schools and municipalities.* Doctoral dissertation University of California Los Angeles.

Fardella, C., & Sisto, V. (2015). Nuevas regulaciones del trabajo docente en Chile. Discurso, subjetividad y resistencia. *Psicologia & Sociedade, 27*(1), 68–79. https://doi.org/10.1590/1807-03102015v27n1p068

León, M. G. (2008). *Calidad docente y rendimiento escolar en Chile.* Tesis para optar al grado de Magíster en Ciencias de la Ingeniería, Pontificia Universidad Católica de Chile. http://repositorio.uc.cl/xmlui/bitstream/handle/123456789/1447/507135.pdf?sequence=1

Law N°19.933. (2004). Diario Oficial de la República de Chile, Santiago, Chile, 12 de febrero de.

Law N°19.961. (2004). Diario Oficial de la República de Chile, Santiago, Chile, 14 de agosto de.

Law N°20.501. (2011). Diario Oficial de la República de Chile, Santiago, Chile, 26 de febrero de.

Law N°20.903. (2016). Diario Oficial de la República de Chile, Santiago, Chile, 1 de abril de.

Manzi, J., Strasser, K., San Martin, E., & Contreras, D. (2008). Quality of education in Chile: Final report of the inter American development bank project. Washington, DC: BID. Recuperado en abril de 2009 de http://www.iadb.org/res/laresnetwork/files/pr300finaldraft.pdf

Ministerio de Educación. (2008). El Marco para la Buena Enseñanza. Santiago, Chile. https://www.docentemas.cl/docs/MBE2008.pdf

Ministerio de Educación. (2019). Marco para la Buena Enseñanza de Educación Parvularia. Santiago, Chile. https://parvularia.mineduc.cl/wp-content/uploads/sites/34/2019/08/MBE_EP-Final.pdf

Roa-Tampe, K. A. (2017). La evaluación docente bajo la óptica del desarrollo profesional: el caso chileno. *Educación y Educadores, 20*(1), 41–61. https://doi.org/10.5294/edu.2017.20.1.3

Roa-Tampe, K. A. (2018). La docencia como profesión: factores contextuales y sociales que influyen el desempeño del rol docente. Tesis para optar al grado de Doctor en Sociología, Pontificia Universidad Católica de Chile. Recuperado en agosto 2020 en https://repositorio.uc.cl/xmlui/bitstream/handle/11534/22091/Karin%20Roa%20Agosto%202019.pdf

Rosales, R. (2018). Percepciones sobre el Sistema de Evaluación Docente y su impacto en la práctica profesional Estudio cualitativo en docentes de establecimientos municipales. Tesis para optar al grado Magíster en Medición y Evaluación de Programas Educacionales, Pontificia Universidad Católica de Chile. https://repositorio.uc.cl/handle/11534/22317

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

Santelices, V., Taut, S., Araya, C., & Manzi, J. (2013). Consecuencias a nivel local de un sistema de evaluación de profesores: el caso de Chile. *Revista Estudios Pedagógicos, 39*(2), 299–328.

Santiago, P., Benavides, F., Danielson, C., Goe, L., & Nusche, D. (2013). Teacher evaluation in Chile. *OECD Reviews of Evaluation and Assessment in Education.* OECD Publishing.

Sepúlveda, A., Hernández, C., Peña, S., Troyano, M., & Opazo, M. (2019). Evaluation of teacher performance in Chile: Perception of poorly evaluated teachers. *Cadernos de Pesquisa 49*(172), 144–163. https://doi.org/10.1590/198053145792

Sisto, V., Montecinos, C., & Ahumada, L. (2013). Disputas de significado e identidad: la construcción local del trabajo docente en el contexto de las políticas de evaluación e incentivo al desempeño en Chile. *Universitas Psychologica, 12*(1), 173–184.

Subsecretaría de Educación Parvularia. (2019). Marco para la Buena Enseñanza de Educación Parvularia. Santiago, Chile: Ministerio de Educación.

Sun, Y., Levy, D., Cortés, O., Ramos, J., & Rojas, M. (2017). ¿Cómo son las percepciones y experiencias de docentes y directivos en torno a la Evaluación Docente? *Midevidencias, 14*, 1–8. http://www.mideuc.cl/wp-content/uploads/2017/MidEvidencias-N14.pdf

Taut, S., Jiménez, D., Puente-Duran, S., Palacios, D., Godoy, M. I., & Manzi, J. (2019). Evaluating the quality of teaching: can there be valid differentiation in the middle of the performance distribution? *School Effectiveness and School Improvement, 30*(3), 328–348. https://doi.org/10.1080/09243453.2018.1510842

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

Taut, S., Santelices, V., & Manzi, J. (2011b). Estudios de validez de la Evaluación Docente. In J. Manzi, R. González, & Y. Sun (Eds.), *La Evaluación Docente en Chile* (pp. 157–175). Santiago: Pontificia Universidad Católica de Chile.

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

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

Taut, S., & Sun Y. (2014). The development and implementation of a national, standards-based, multi- method teacher performance assessment system in Chile. *Education Policy Analysis Archives, 22*(71), 1–30.

Taut, S., Valencia, E., & Escobar, J. (2012b). La validez de la Evaluación Docente en Chile usando como criterio estimaciones de Valor Agregado de profesores de Enseñanza Media. Informe Técnico Mide UC. http://mideuc.cl/wp-content/uploads/2013/01/IT1202.pdf

Taut, S., Valencia, E., Palacios, D., Santelices, V., Jiménez, D., & Manzi, J. (2014). Teacher performance and student learning: Linking evidence from two national assessment programs. *Assessment in Education: Principles, Policy y Practice, 23*(1), 53–76. https://doi.org/10.1080/0969594X.2014.961406

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; *Program. Evaluation and Program Planning, 33*, 477–489. https://doi.org/10.1016/j.evalprogplan.2010.01.002

Torres, D.,& Zapata, A. S. (2019). Portafolio en la evaluación docente en Chile: Recolección de evidencia de validez como parte del proceso de construcción del instrumento. In J. Manzi, M. R. García, & S. Taut (Eds.) Validez de evaluaciones educacionales en Chile y Latinoamérica (pp. 605–638). Santiago, Chile: Ediciones UC.

Universidad Técnica Metropolitana (2017). Condiciones de instalación de la Ley N°20.501 de Calidad y Equidad de la Educación. Informe final. Recuperado de https://documentos.serviciocivil.cl/actas/dnsc/documentService/downloadWs?uuid=1e5835c2-e0d1-4d0d-b2b3-ad8ba9a4aa67

Urriola, K. (2013). Sistema de evaluación del desempeño profesional docente aplicado en Chile. Percepciones y vivencias de los implicados en el proceso. El caso de la ciudad de Concepción. Tesis para optar al grado de Doctor, Universitat de Barcelona - España. http://diposit.ub.edu/dspace/bitstream/2445/50737/1/01.KMUL_1de2.pdf

Chapter 9
Teacher Evaluation in Mexico

Sylvia Schmelkes

Abstract Teacher evaluation has acquired vast relevance at the international level. There are at least two reasons that explain it. First, it focuses on teaching practice, which is the factor closest to the student and therefore to learning. Second, it represents an important input that influences initial training, keeping teachers up to date and providing a system of professional improvement. The Mexican case is inspired by these same convictions even though its history of teacher evaluation is recent. This chapter addresses the *Carrera Magisterial* (Teaching Career Program) and the *Sistema Profesional Docente* (SPD, Professional Teacher System), as well as the political and social factors that promoted its implementation, such as the educational reform of 2013 and 2019, and the creation of *the Instituto Nacional para la Evaluación de la Educación* (National Institute for Educational Evaluation) in 2002. In addition, the challenges, opportunities, and results derived from the Mexican experience are discussed.

9.1 Background on the Evaluation of Teachers in Mexico

Teacher evaluation in Mexico has a relatively short history. For many years, teachers received salary increases through a system called *the Escalafón Vertical* (Vertical Promotion System)**.** As in many other countries, Mexico adopted a scale that measured years in the system, teacher preparation, and evaluations not based on standardized instruments, but rather, in many cases, granted in a discretional manner on the part of immediate supervisors. In the case of Mexico, this scale started being used in 1930 and was reformed for the last time in 1973 with the publication of the "*Reglamento de Escalafón de los Trabajadores al Servicio de la Secretaría de Educación Pública*" (Regulations for the Promotion of Workers in the Service of the Ministry of Public Education). The promotion scheme assessed knowledge

S. Schmelkes (✉)
Instituto de Investigaciones para el Desarrollo de la Educación, Universidad Iberoamericana
Ciudad de México, Prol. Paseo de la Reforma 880, 01219 Ciudad de México, México
e-mail: sylvia.schmelkes@ibero.mx

J. Manzi et al. (eds.), *Teacher Evaluation Around the World*,
Teacher Education, Learning Innovation and Accountability,
https://doi.org/10.1007/978-3-031-13639-9_9

(academic degree plus professional and personal improvement), aptitude (efficiency and initiative), seniority (years of service), and discipline and punctuality (Guzmán Marín, 2018). However, as Martínez Rizo (2016) has indicated, worldwide this promotion strategy has not proven its ability to distinguish between good and bad teachers. It has lent itself to arbitrary decisions or, in other cases, to a zero-level demand, so that it was enough for teachers to show up to work daily to obtain salary increments. In the worst cases, among which is the case of Mexico, even serious offenses on the part of teachers, such as customary absences, alcoholism, or even more serious crimes, were not grounds for dismissal.

9.1.1 The Teaching Career Program

The limitations of the promotion scheme led to the search for more objective methods for evaluating and promoting teachers. In 1993, Mexico established the *Teaching Career Program*, a system of salary increments for teachers and school leadership positions according to their performance and that of the students, as a consequence of the *Acuerdo Nacional para la Modernización de la Educación Básica* (National Agreement for the Modernization of Basic Education), signed by the federal government, the 31 state governments and the *Sindicato Nacional de Trabajadores de la Educación* (SNTE, National Union of Education Workers). It was one of four fundamental changes in educational policy, together with the decentralization of basic and teacher education to the federal entities, the initiative for fundamental curricular reform, and the beginning of a policy of social participation in education. *The Teaching Career Program* was the consequence of the intent to revalue the role of the teacher and establish an effective system for keeping the profession up to date. It was the first attempt in Mexico to link teacher salaries to their training and performance (Echávarri & Peraza, 2017; Gluyas & González, 2014). Its primary objective was "to stimulate the quality of education and establish a clear means of professional and material improvement as well as the social conditions of the teacher" (*Diario Oficial de la Federación* [DOF] Official Gazette of the Federation, 1992, p. 13, cited in Plá, 2019). It is important to highlight the voluntary nature of the Program (Guzmán Marín, 2018).

The governing body of the *Teaching Career Program* was a National Commission SEP[1]-SNTE with its related commissions in the states, also composed by members from government and from the teachers' union. In each educational institution, an Evaluation Commission responsible for disseminating and operating the program locally was established. The program was expected to shape a career of professional development. It was structured in a five-stage path (A, B, C, D, and E), with the aim of improving the quality of teaching and, at the same time, allowing teachers to move up to positions within the service while they remained in schools, rather than

[1] *Secretaría de Educación Pública* (SEP, Ministry of Public Education).

being commissioned for managerial positions within the local or federal ministries of education or the teachers' union.

At the beginning of the *Teaching Career Program,* the evaluation of the teachers considered the following elements: (a) years of experience, (b) teacher professional development and education (consisting of "coursework update modules" and teaching degrees), (c) a peer review, and (d) student performance (Echávarri & Peraza, 2017; Martínez Rizo & Blanco, 2010). The corresponding salary stimulus was established beginning with an additional 25% at level A and continuing up to 200% at level E (Ducoing, 2019). From the start of this program, a permanent student assessment program was set up in which the teachers also participated, and thus, a systematic policy for verifying the performance of both the teacher and the students was set in motion (Ducoing, 2019). However, the results of student assessments were not made known until 2010, and when they were published by the *Dirección General de Evaluación* (General Directorate of Evaluation), they barely reached the schools and the organizations that could use them to make informed decisions (Fernández & Midaglia, cited in Martínez Rizo & Blanco, 2010).

The *Teaching Career Program* went through several stages and the relative weight of the results of the tests applied to the students varied in the teacher evaluation with each new incarnation. The program was evaluated by Santibáñez et al. (2006), who showed the limited relationship of each of the factors considered with the level reached by teachers in the scheme by stages, as well as with the achievement of their students. The factors considered were also unrelated to the teachers' results in the professional preparation tests. Santibáñez et al. (2006) conclude that the *Teaching Career Program* responded more to the need to compensate for poor teacher salaries that occurred as a consequence of the so-called lost decade[2] than to a need to improve the quality of education in the country. Ducoing (2019) found that there was a negative effect on student test scores after primary and secondary teachers received the salary stimulus. In other words, once the teachers were incorporated into the program, a certain decrease was observed in student scores, which got worse as students were promoted to subsequent levels (Ducoing, 2019).

The year 2000 was an important year in the history of Mexico because for the first time in 71 years a president from a different party than the long-ruling PRI was elected.[3] In the process of putting together the new government program, the "transition team" recommended the creation of a technical body, with a considerable degree of autonomy, that would be in charge of carrying out large-scale evaluations of the educational system, until then the exclusively responsibility of the Ministry of Public Education. The *Instituto Nacional para la Evaluación de la Educación* (INEE, National Institute for Educational Evaluation) was thus created by decree in 2002. With its creation, two important changes occurred: (1) although the INEE was

[2] The "lost decade" refers to the eighties of the last century, which in Latin America were marked by hyperinflation and a consequent recession that impeded the economic growth of practically all the countries of the region for a prolonged period.

[3] For the first time in 71 years, a candidate came to power who did not belong to the *Partido Revolucionario Institucional* (PRI, Institutional Revolutionary Party), which had become the hegemonic party in Mexico after the revolution.

created depending on the Ministry, without full autonomy, it was granted technical autonomy and the Ministry of Public Education ceased to be judge and jury in its evaluations; and (2) the evaluations became public by law. During its early years, the INEE focused most of its efforts on the design of standardized tests—the Educational Quality and Achievement Tests (EXCALE)—to track achievement of students in basic education, as well as on the application of international assessment tests: the Programme for International Student Assessment (PISA) and the test of the *Laboratorio Latinoamericano de Evaluación de la Calidad de la Educación* (LLECE, Latin American Laboratory for Evaluation of the Quality of Education) of the UNESCO Regional Office for Latin America and the Caribbean.

The results of national and international tests revealed two serious problems in the national educational system: the low performance of students—around half of them below the basic level considered necessary to continue studying or to be able to face the demands of the current society—and the enormous inequality depending on the type of school attended, the locality in which one lived, the degree of marginalization of the area in which the school was located, and the parents' income and schooling. In the social imaginary, these results, especially those referring to the low performance of students on standardized tests, were related to the performance of teachers. As indicated by Ibarrola (2018), there was ample evidence about the precarious material conditions of schools and the poor socioeconomic conditions of many students and even many teachers, but the emphasis was wrongly placed on the poor performance of the latter.

9.1.2 Alliance for Quality in Education

This explains that, with the *Alianza por la Calidad de la Educación* (Alliance for Quality in Education), an agreement signed by the Secretariat of Public Education and the teachers union (SNTE) in 2008, it was agreed that admission to teaching would no longer be decided by the SNTE but would be based on an objective evaluation that as of 2009 would be applied by an independent body (Independent Federalist Evaluation Body) (Flamand et al., 2020; Martínez Rizo & Blanco, 2010). The teachers union (SNTE) also participated in that body. The evaluation consisted of an 80-item test that comprised three substantive areas: teaching content, didactic skills, and basic intellectual skills. Although the test was applied in 2008 and 2009, it was not universal and was widely questioned, and its history was brief.

Another product of the *Alliance for the Quality of Education* was the *Programa de Estímulos a la Calidad Docente* (Program of Incentives for Teaching Quality). Participation in this program was also voluntary, and it recognized both individually and collectively the teachers whose students obtained the best learning achievements on the *Evaluación Nacional de los Logros Académicos en Centros Escolares* (ENLACE, National Assessment of Academic Achievement in School Centers), a universal test also a product of the Alliance, which was applied to all students from third grade onwards, every year, by teachers from the same school who are not from the evaluated group. Linking the evaluation of the students to economic stimuli for the teachers generated many perverse effects that include teaching to the test, dissuading the attendance of students with learning deficits on the day of the exam, not accepting students with special educational needs or speakers of a language other than Spanish, and even illicitly trafficking the test prior to its application (Backhoff & Contreras, 2014).

9.1.3 The Role of the National Union of Education Workers (SNTE)

As can be seen from this brief historical account of the background of teacher evaluation in Mexico, the role of the SNTE has been central. The origin of its importance in the professional career of teachers is historical and is widely described in the now classic book by Arnaut (1993). In a schematic way, it can be said that the expansion of the Mexican educational system from the creation of the Ministry of Public Education one hundred years ago (1921), and very notably from the since 1934 with socialist education, allowed the Mexican State to distribute its representatives among the teachers and increasingly throughout a good part of the Mexican Republic. As such, they were called upon to fulfill functions of a diverse nature, with one fundamental role being mainly political, that of guiding the population toward voting for the party in power. In exchange for this important function, the government granted the SNTE a series of prerogatives, among which the control of teaching positions, school changes, the use of salary scales and vertical promotions to supervision and school leadership positions were the most important. For decision-making regarding the location and mobility of teachers, joint commissions were officially established in which the Secretariat of Public Education and the SNTE participated, each with 50% of the votes. The political power that the SNTE acquired as a consequence was enormous.

Another fundamental concession was the mandatory affiliation of all teachers and education administrators to SNTE, which in addition to leading to it eventually becoming the largest union in Latin America, gave it great economic power. Teaching positions and changes both geographically and at the level of vertical promotion were handled, in the absence of objective evaluation mechanisms, as favors from the SNTE to its bases. This level of power soon led to the corruption of a significant number of union leaders who offered such favors in exchange for substantial payments that enriched them personally. Political and economic power and control over teachers allowed union positions to be used on many occasions as a means of access to important political or elected positions. In the SNTE, there was little transparency, and leaders exercised authoritarian power over their affiliates.

The tight union control over the teachers, together with the total absence of democratic procedures to elect their leaders, led to dissident movements that brewed from within. The most important of these, the *Coordinadora Nacional de Trabajadores* *de la Educación* (CNTE, National Coordinator of Education Workers), emerged in 1979 and defined its struggle as its intent to democratize the teachers union. This movement acquired strength, especially in the poorest states in the southeast of the country, and notably in Oaxaca, Chiapas, and Guerrero, although it is represented in practically all of the 32 states. The case of Oaxaca is emblematic, since Section 22 of the Union was fully occupied by the CNTE, which to date controls the State Institute of Public Education of Oaxaca and with it all the educational decisions that a state can make. In fact, the dissidence represented by the CNTE has become more of a faction that fights for union power and control than for a democratic approach within the SNTE. Moreover, its operating methods when it has political and educational power, as in Oaxaca, are akin to those of the SNTE. The consequence of the success of the CNTE's struggle has been for it to gain control over teaching posts and movements in place of the SNTE. However, the SNTE is a powerful and complex structure that has a presence, through a teacher representative, in each and every one of the country's schools and its ability to communicate with its members is powerful.

This history of teacher unionism in Mexico is essential to understand the outcome of the educational reform carried out between 2013 and 2019, one of whose fundamental elements was teacher evaluation.

9.2 The Educational Reform 2013–2019

In 2013, with the PRI regaining political power, the so-called Pact for Mexico was launched. The three main political parties came together to pass structural reforms in Congress that were considered fundamental to turn around and strengthen development in the country. One of these was an educational reform that sought to improve the quality of education. This reform was proposed to increase the quality of basic education, increase enrollment and improve the quality in the high school and higher education systems. For this, it was considered essential "for the State to regain control of the national educational system" (Bracho & Zorrilla, 2015) in direct reference to the need to reduce the power of the SNTE. The Pact for Mexico made possible the reform of Article 3 of the Political Constitution of the United Mexican States, which refers to education. Just the day before this change was approved, the government arrested the SNTE life-long leader, Elba Ester Gordillo, on charges of tax fraud. An attempt was made to send a clear message to the SNTE regarding the intention of the State to "regain control" of education.

Three fundamental changes were proposed in Article 3 of the Constitution (Official Gazette of the Federation [DOF], 2013):

(1) The definition of the quality of education. "The State will guarantee the quality of compulsory education so that educational materials and methods, school organization, educational infrastructure and the suitability of teachers and the school leadership's management guarantee the maximum achievement of student learning."

(2) The creation of the Professional Teaching Service and the determination that entry into the teaching service and promotion to positions with managerial or supervisory functions in basic and high school education provided by the State, "will be carried out through public exams that guarantee the suitability of the corresponding knowledge and skills. The regulatory law will establish the criteria, terms, and conditions of the mandatory evaluation for entry, promotion, recognition, and permanence in the professional service with full respect for the constitutional rights of education workers. All entries and promotions that are not granted in accordance with the law will be void."

(3) The creation of the National Educational Evaluation System and the definition of the National Institute for the Evaluation of Education (INEE) as "an autonomous public body, which functions as a legal entity with its own assets. The INEE's objective was to evaluate the quality, performance, and results of the National Education System in educational levels of preschool, primary, middle school, and high school. To do this, it should (a) design and carry out the measurements that correspond to components, processes or results of the system; (b) issue the guidelines to which the federal and local educational authorities will be subject to carry out the evaluation functions that correspond to them, and (c) generate and disseminate information and, based on this, issue guidelines that are relevant to improve the quality of education and its equity, as an essential factor in the search for social equality" (INEE, 2015a: 41).

To carry out the teacher evaluation established by the Reform of Article 3 of the Constitution, a complex institutional arrangement was set up in which the following bodies participated: INEE as a regulatory and supervisory entity; the federal educational authority through the National Coordination of the Professional Teaching Service as the entity responsible for the design of the evaluation instruments and the organization of the evaluations, and the state educational authorities as responsible for their application (see Fig. 9.1). This complex organizational fabric from its initial definition explains an important part of the difficulties in the implementation of the teacher evaluation that we will analyze in what follows.

It is important to note that the evaluation of teaching performance was conceived in the same constitutional reform, as an input that should serve the following:

> To give greater relevance and capabilities to the national system of education, updating, training and providing professional development for teachers within the framework of the creation of a professional teaching service. The evaluation of teachers must have as its first purpose, that they and the educational system have well-founded references for reflection and dialogue leading to better professional practice. The educational system must provide the necessary support so that as a priority, teachers can develop their strengths and overcome their weaknesses. (from the Constitution of the United Mexican States)

As can be seen, the evaluation of teacher performance was never conceived as punitive; from the beginning, its function was defined as formative.

The Constitutional Reform led to the issuance of three secondary laws: the reformed General Law of Education, the General Law of the Professional Teaching Service, and the Law of the National Institute for the Evaluation of Education, now as an autonomous body, which were approved in September 2013.

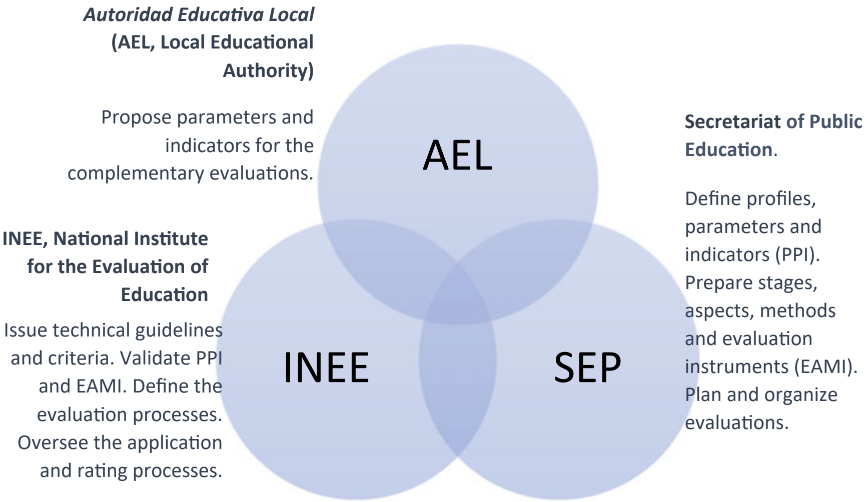

Autoridad Educativa Local (AEL, Local Educational Authority)
Propose parameters and indicators for the complementary evaluations.
AEL
INEE, National Institute for the Evaluation of Education
Issue technical guidelines and criteria. Validate PPI and EAMI. Define the evaluation processes. Oversee the application and rating processes.
INEE
SEP
Secretariat of Public Education.
Define profiles, parameters and indicators (PPI). Prepare stages, aspects, methods and evaluation instruments (EAMI). Plan and organize evaluations.

Fig. 9.1 Organizational framework of the teacher evaluation established in the reform of Art. 3° of the Constitution. *Notes* Adapted from *Evaluation of Teacher Performance Model 2017* (p. 2), INEE (2017). https://local.inee.edu.mx/w.-content/uploads/2019/01/diptico-dic17.pdf

9.3 The Teacher Evaluation 2013–2018

9.3.1 Evaluation for Admission to Teaching

The *Ley General del Servicio Profesional Docente* (LGSPD, General Law of the Professional Teaching Service, 2013) stipulated in one of its articles that, "the Institute, the Secretariat, the local educational authorities and the Decentralized Organizations must carry out during the month of July 2014 an exam… for entry to the Service in Basic and Higher Secondary Education." The evaluations had to be ready ten months after the Law was approved, and thus, it was necessary to work in a hurry to comply with this provision. To achieve this, a third actor was incorporated into the complex network of bodies responsible for teacher evaluation: the *Centro Nacional de Evaluación de la Educación Superior* (CENEVAL, National Center for the Evaluation of Higher Education), which was commissioned by the General Coordination of the Professional Teaching Service to prepare the instruments for entry examinations.

The teaching entrance evaluation consisted of two stages: The first was an exam that measured the curricular or disciplinary knowledge of the level and the subject to be taught, as well as pedagogical knowledge, called "knowledge and skills for professional practice" (100 items). The second consisted of an examination of "intellectual skills and ethical-professional responsibilities" (100 items). For certain aspiring teachers (of arts, of language in the case of aspiring indigenous teachers, of the State-approved subject matter and of technology), there were also additional or complementary examinations. The wide diversity of types of teachers in the country implied a huge effort to design around 27 different instruments according to the level and subject taught.

In March 2014, the first round of evaluations for teaching entry and promotion to managerial positions in basic and high school education was called, which was held in July of that year (see Fig. 9.2). It was a massive endeavor with 149,978 aspiring elementary education teachers, 42,776 aspiring high school teachers, and 1165 aspiring teachers for director positions responding online for exams in application centers. Given the rush with which it was prepared, the first experience was fraught with difficulties. The teachers' union, which did not welcome the educational reform, took advantage of the many mistakes made in the first experience to revile it. The INEE, as the body responsible for regulating the evaluation and supervising it, requested the UNESCO Institute for Educational Planning in Buenos Aires to carry out an external evaluation of this experience, which also involved witnessing its application (Fumagali & López, 2015) to suggest ways to improve its implementation in the future. Its conclusions and recommendations are highlighted here:

- In general, the process was well valued because it was faithful to the design and guidelines issued by the INEE.
- Those involved had a positive perception of the evaluation experience
- Technical and organizational successes were observed.

Despite the above, issues that caused discomfort in the teachers were identified (the excessive length, the low readability of the texts, some errors in the questions, delay in the administration of the exam, among others). Many problems arose related to inadequate communication between the different actors involved in administering the test and the consequent lack of ownership of the processes, especially by those who were in charge of the exam sites.

Some more serious aspects also appeared that were the object of strong criticism: the fact that the operation was militarized—for security reasons, it was decided that the army would monitor it; the fact that so little room for action was left to the state educational authorities, which strengthened the perception of excessive central control of the process; and the discomfort of the SNTE that its members were not allowed to participate as observers—the observers were members of civil society organizations and parents. The recommendations resulting from this international evaluation, as well as those derived from the process of supervision of the application carried out by INEE, were incorporated in the successive annual administration of the entry exam and evaluations for promotions for teachers already in the system.

The results of the teaching entry evaluation improved over the years, as shown in Table 9.1. However, the differences between the state entities were important. Thus, for example, in administering the exam corresponding to the 2017–2018 school year, when 59% of applicants nationwide obtained results that defined them as suitable

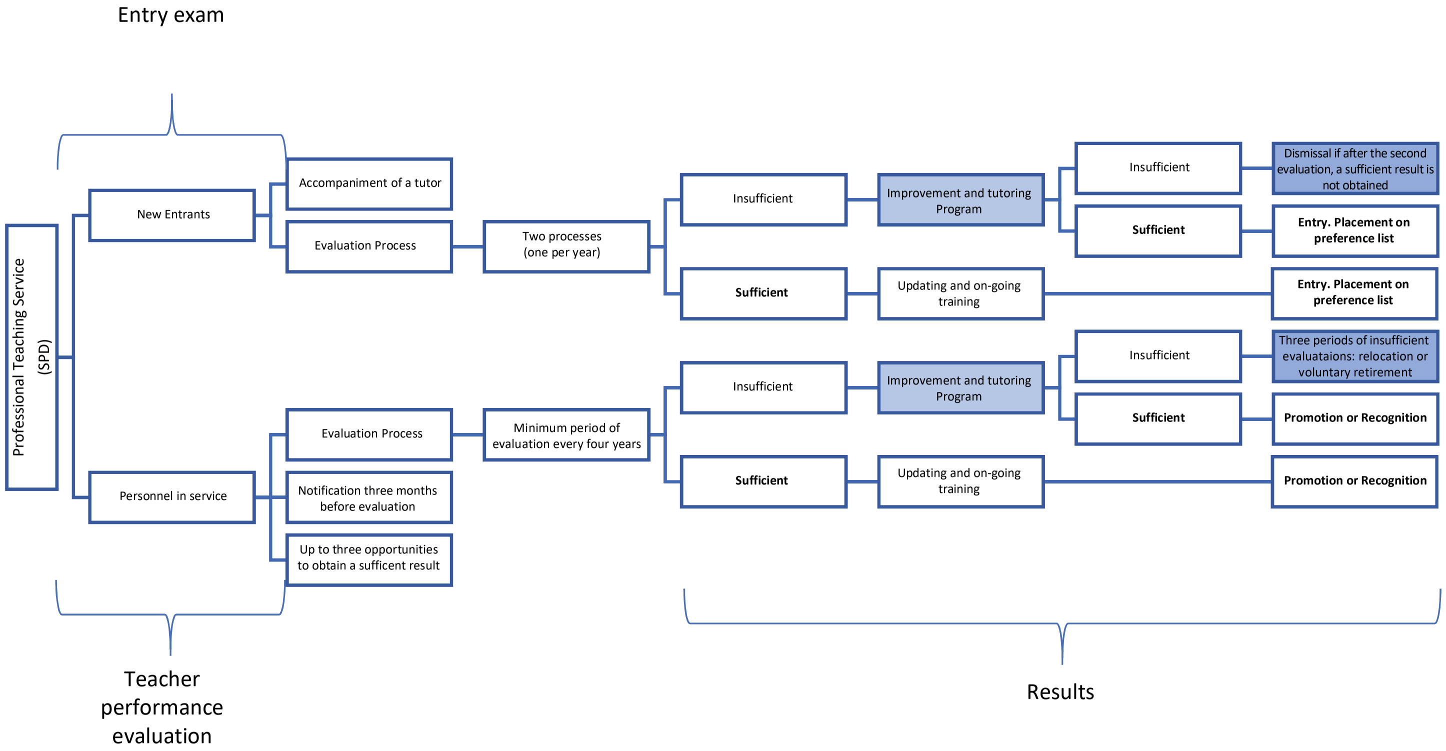

Entry exam
Professional Teaching Service (SPD)
New Entrants
Accompaniment of a tutor
Evaluation Process
Two processes (one per year)
Insufficient
Improvement and tutoring Program
Insufficient
Dismissal if after the second evaluation, a sufficient result is not obtained
Sufficient
Entry. Placement on preference list
Sufficient
Updating and on-going training
Entry. Placement on preference list
Personnel in service
Evaluation Process
Notification three months before evaluation
Up to three opportunities to obtain a sufficent result
Minimum period of evaluation every four years
Insufficient
Improvement and tutoring Program
Insufficient
Three periods of insufficient evaluataions: relocation or voluntary retirement
Sufficient
Promotion or Recognition
Sufficient
Updating and on-going training
Promotion or Recognition
Teacher performance evaluation
Results

Fig. 9.2 Evaluation process of the professional teaching service. *Notes* Adapted from *2014 Annual Program of the evaluation processes of the Professional Teaching Service* (p. 35). http://servicioprofesionaldocente.sep.gob.mx/portal-docente-2014-2018/content/general/docs/normatividad/PROGRAMA_ANUAL_2014%20_PROCESOS_EVALUACION.pdf

Table 9.1 Teacher applicants evaluated by year of application and percentage with results that defined them as suitable for teaching in basic education

Year of application (school year)	Applicants evaluated	% of applicants with "suitable" results
2014–2015	130,503	39
2015–2016	116,036	52
2016–2017	108,317	60
2017–2018	120,565	59
2018–2019	132,450	60

Notes Adapted from *Supervision report on the evaluation of teacher performance in basic and high school education, in the 2017–2018 school year* (p. 11), INEE (2018a, 2018b, 2018c). https://historico.mejoredu.gob.mx/wp-content/uploads/2019/04/P1F229.pdf; y *La Educación Obligatoria en México. 2019 Report* (p. 72), INEE (2019a, 2019b). https://www.inee.edu.mx/wp-content/uploads/2019/04/P1I245.pdf

(sufficient) for teaching, the entity that registered the highest percentage in this category was Querétaro with 73% followed by Baja California with 72%; meanwhile, Tabasco and Michoacán registered the lowest figures, 35% and 42%, respectively. Two of the poorest states, Chiapas and Guerrero. Did not present data (INEE, 2019a).

The results of the performance evaluations are grouped as follows: Level I refers to insufficient; Level II, sufficient and organized command of knowledge and skills; Level III, in addition to showing a sufficient and organized domain of knowledge and skills, the applicant demonstrates a broad capacity to implement them didactically. Once the second performance evaluation was carried out, the categories corresponding to the groups were the following: A, in both exams the applicant obtained Level III; B, in an exam the applicant obtained Level III and in the other Level II; C, in the two exams the applicant obtained Level II (DOF, 2017).

Those with the best results in the 2014 admission evaluation obtained the highest percentage with results "Sufficient A" (22.8%) and "Sufficient B" (66%) in the performance evaluation at the end of the second year. The association is positive and significant equally for teachers of basic and higher secondary education, although it was higher for the first[4] (INEE, 2018a).

After four years in effect, the entrance evaluation and the evaluation for promotions were becoming established among the teachers and achieving good acceptance. The satisfaction surveys applied to aspiring teachers also show an improvement in the processes prior to and during the evaluation throughout the three years in which they were applied (INEE, 2018a, 2018b, 2018c). It was deemed appropriate that merit defined who became a teacher and who had priority to choose the workplace among the available vacancies. People recalled how in the past it was necessary to purchase access to teaching and leadership positions and various other types of favors requested by some union leaders. The satisfaction survey corresponding to the 2017–2018 evaluation shows percentages higher than 80% of high satisfaction in

[4] (tau-b: approximate $T = 33.905$, $p < 0.0001$) and (tau-c: T approximate $= 19.639$, $p < 0.0001$).

the dimensions related to the processes prior to the administration of the exam, with the exception of the bibliography and the study guides (71% are highly satisfied) and were similar in the dimensions relative to the process involved in administering the exam (the lowest percentage refers to the exam, 67%). The percentages of high satisfaction drop considerably, to just over 43%, in the dimensions related to the post-evaluation stage, that is, in the consequences of the evaluation on the allocation of sites of schools in which teachers were to work (INEE, 2018b).

Despite the above, the SNTE was not happy. They never protested openly and even collaborated in dissemination of information about the upcoming exams. But the entrance exam effectively detracted from a source of power and eliminated a wide space of corruption consolidated for decades. This subterranean unease manifested itself, as will be seen later, in a distortion of the performance evaluation's purpose in order to generate animosity on the part of the teachers, most visibly toward the educational reform of 2013. The CNTE, for its part, simply did not accept the evaluation. In the entities in which it had control, it did everything possible to prevent the participation of applicants in the evaluation—applications had to be submitted in neighboring states—and, later, it harassed those who obtained a place through the exam process. It prohibited the teachers affiliated with its sections from participating in the evaluations for promotion to managerial positions. Throughout the 4 years in which the teacher evaluation was applied, they sowed fierce opposition to the educational reform and in many cases achieved the support of the communities in which this section was present.

There is little information about the impact of the teaching entrance evaluation. The little that is available is encouraging. De Hoyos and Estrada (2018) found a high correlation between the scores in the ENLACE test of higher secondary Education and the probability of obtaining results as "suitable" in the *Concurso de Ingreso del Servicio Profesional Docente* (Entry Exam for the Professional Teaching Service). Those who were selected to practice teaching belonged to the higher performance percentiles in the last year of higher secondary according to the Enlace test compared to those who entered teaching before there was a universal admission evaluation (see Table 9.2). These results demonstrate the ability to select the evaluation applicants from the best students, judging by the results on standardized tests, and also show that the teachers who entered teaching through the exam process were better students, at least in higher secondary, than those who entered before screening for entry was universal.

The most interesting data, however, is that those who entered teaching by competition, in 99.5% of the cases, were approved in the performance evaluation at the end of the second year (INEE, 2015b). The entrance evaluation and the performance evaluation do not measure the same thing: The second evaluation tries to approach what the teacher does in the classroom. This result, therefore, would seem to indicate that, as the article by De Hoyos and Estrada (2018) and its title suggest, the teachers who start to work in the classrooms of Mexico are better teachers as a consequence of the entrance evaluation.

Table 9.2 Percentile on the ENLACE test of secondary education of those who entered teaching before and after the universalization of the entry exam

Moment	Year	Percentile
Before the universalization of the teaching entry exam	2012	58.1
	2013	56.6
After the universalization of the teaching entry exam	2015	62.1
	2016	60.9

Notes Adapted from "Did the teachers get better? Yes" (*Los docentes mejoraron? Si*!), De Hoyos and Estrada (2018), *Nexos*. https://www.nexos.com.mx/?p=39531

9.3.2 Teacher Performance Evaluation

The teacher performance evaluation, mandated by the General Law of the Professional Teaching Service, began one year after the teaching entry exam, in 2015. As we have already pointed out, the purpose of this evaluation was essentially formative: to provide feedback to the teacher about his/her areas of improvement and to provide feedback to the educational system in relation to the design of initial training and ongoing professional development. Despite this, the General Law of the Professional Teaching Service did stipulate that teachers would have three opportunities to take the evaluation and would have to leave the system if passing results were not achieved by the third attempt. The law also defined the performance evaluation as mandatory and stated that those who were summoned and did not appear for the evaluation would be required to leave the system. These two sanctions are those that led teachers to identify the performance evaluation and the educational reform, as "punitive," an adjective that penetrated the perception and spirit of teachers increasingly throughout the period in which the reform was in effect and, in the end, was the fundamental cause for its repeal.

As with the teaching entry evaluation, the performance evaluation had to be hastily designed. Also, for the performance evaluation, the National Coordination of the Professional Teaching Service turned to CENEVAL for the elaboration of the instruments. An additional actor was incorporated, the *Instituto Latinoamericano para la Comunicación Educativa* (ILCE, Latin American Institute for Educational Communication), which was entrusted with managing the performance evaluation platform. INEE had the function of regulating and supervising the entire process. The first performance evaluation was applied in 2015, and 152,000 teachers were called to participate. There were 132,000 who responded to the first call. The open opposition to its implementation by—above all, but not only—the CNTE, led the Ministry of Public Education to have the testing sites guarded, resorting to the police who, unarmed but in spectacular operations, were in charge of monitoring the sites and in some cases transferring teachers to them. This generated enormous discomfort even among those teachers who did not object to the evaluation, as well as in broad sectors of society in general.

The first version of the performance evaluation consisted of four "stages." The first was a report on the fulfillment of professional responsibilities, prepared by the direct superior of the teacher or manager and uploaded to the platform. The second stage was a teaching evidence file, which was evaluated by means of a rubric. The third stage consisted of an examination of knowledge and didactic skills that favored student learning using 77 different instruments, one for each different type of teacher, and the fourth consisted of a justified didactic plan that was also scored with a rubric. The first two stages were carried out by the teacher using the online platform; the third and fourth were carried out at the application sites. As can be seen, a multidimensional evaluation was designed, both quantitative and qualitative, which sought to get as close as possible, albeit indirectly, to evaluating not only the knowledge necessary to teach, but also the teaching practice itself. The result was a complex design.

As on the occasion of the first teaching entry evaluation, INEE requested an external evaluation of the process, in this case from UNESCO's Regional Office for Latin America and the Caribbean (INEE, 2017). The study was rich in details. It pointed out important weaknesses in the process that, from the evaluators' point of view, put the sustainability of the evaluation at risk. In preparing the evaluation, the main problem was in the elaboration of evaluation instruments: item banks that were partially revised and multiple-choice items that included replacement questions for those that did not meet the standards, since piloting the test *ex ante* was not possible. This forced the number of questions on the tests to be significantly increased and, consequently, increased the examination time"[5] (INEE, 2017, p. 22). In the preparation activities for the evaluation (selection, information, registration and support) there were many failures that were directly perceived by the teachers and generated a lot of discomfort. In the application, the main annoyances of the teachers were reported in the stages carried out at the test sites: their location, in many cases far from their homes; the admission processes; the mistreatment by the CENEVAL proctors during the application, and the 8 hours that they had to spend in front of the computer answering the exam questions. As we have already mentioned, the strong police protection at the test sites with a significant CNTE presence protesting outside some application sites also caused severe annoyances for what was considered by many to be improper treatment (INEE, 2017).

The UNESCO-OREALC evaluation (INEE, 2017) warned very clearly about the risk of not achieving the perception of the evaluation's legitimacy on the part of teachers and society. They placed the complexity of the evaluation in the two agendas of the educational reform: the recovery of the State's control over educational matters and the professionalization of teaching. This explains the distributed governance in the instances involved in the evaluation to which we have already referred and the difficulty coordinating between them in this first evaluation exercise. But the external evaluators perceived from this first moment that the criticism from opposing teachers was seen a criticism of the evaluation process (which did have important flaws), when

[5] A subsequent analysis carried out by the INEE found that only 4 of the 77 instruments had to be eliminated for not complying with the established standards.

in reality what was in dispute was the State's management over education. Since then, serious communication problems have been observed regarding the intentionality of teacher professionalization (INEE, 2017).

Despite the above, the OREALC UNESCO study found that many of the teachers interviewed reported how the pressure of this evaluation led them to study on their own and with their colleagues, which even allowed them to learn about new topics, such as the rules and laws related to their profession. The data collected through a questionnaire answered by a national sample of teachers carried out in November corroborates these reports: fifty-six percentage of the teachers surveyed consider that the evaluation process helped them to learn useful knowledge and skills for the development of their teaching practice (INEE, 2017).

The results obtained by the teachers in the first application of the performance evaluation were generally good: In 2015, only 13.8% of the basic school teachers and 17.3% of the higher secondary teachers were identified at the "Insufficient" level. In 2016, the proportion was even lower, both in basic education (5.6%) and in higher secondary (5.9%). If the results are compared by instrument, in 2016, improvements are observed in educational project both in basic education and in higher secondary. In the case of higher secondary teachers, the results in the disciplinary exams increased the proportion of the highest result by just over 12% and in the didactic knowledge and skills exam by up to 18 percentage points (INEE, 2018a).

The OREALC UNESCO study, together with the supervision, evaluation and analysis efforts carried out by INEE of this first teacher evaluation experience—supervision reports, satisfaction surveys, systematization of complaints and focus groups—led to a reformulation of the evaluation in 2017. This new evaluation model recuperates what worked well in the previous one, eliminates what a careful analysis discovers did not work well, and rethinks the performance evaluation so that it takes place in the school and is linked to context and to the improvement path teachers had to design. It deepens contextualization by referring to the characteristics of the environment and the student group, is more pertinent to better serve teaching practice, emphasizes the formative role of evaluation—offers training before, during and after the process—and makes the evaluation process more accessible. The redesign now consists of three stages: a report on the fulfillment of professional responsibilities, which includes a self-assessment; a teaching project that starts from the diagnosis of the context and the student group and responds to its characteristics; and an examination of curricular and disciplinary knowledge. Stages 2 and 3 are supported by knowledge reinforcement processes and with a support course to develop the teaching project. Only the third stage takes place at a testing site. This reformulation addresses some of the design problems of the previous model, while emphasizing its formative intention. It was applied twice, with much greater acceptance, as shown by the results of the comparison of the satisfaction surveys in the years 2015, 2016, and 2017 (INEE, 2018c). Table 9.3 shows this comparison with the survey applied to teachers after the first performance evaluation. As can be seen, the satisfaction rates increase considerably over the three years, with the exception of the indicator corresponding to the operation of the platform to upload the documents for Stages 1 and 2. The performance evaluation, judging by the perception of the teachers,

Table 9.3 Comparison of the percentage of satisfaction of teachers evaluated in different areas surveyed

Indicator	2015	2016	2017
Advanced notice you were given	49	52	83
Duration of the exam	27	56	84
Total number of exam questions	58	73	81
Length of the exam questions	27	41	61
Professional responsibilities report	58	70	83
Aspects evaluated on the exam	30	56	59
Precision in the wording of the questions	27	48	52
Attention given by the exam proctor	80	89	95
Operation of the computer equipment	76	85	94
Infrastructure of the test environment (classrooms, cafeteria, restrooms)	70	81	91
The relationship of the exam guide and the bibliography with the content of the exam	24	56	56
The operation of the technological platform for Stages 1 and 2	62	75	57

Notes Adapted from *Evaluación de Desempeño* (Results from a Survey of Satisfaction) p. 52, INEE (2018a, 2018b, 2018c). https://www.inee.edu.mx/wp-content/uploads/2018/12/(P1F225.pdf

improved over time, and along these improvements came an increase in its degree of acceptance.

There is little information on the impact of the evaluation carried out on more than 400,000 teachers in the national education system between 2015 and 2018. Weiss et al. (2019) carried out a qualitative study of the teaching practice of a sample of 24 teachers with and without experience who obtained qualifications of outstanding, good, insufficient and sufficient, in the case of novice teachers. The study finds important discrepancies in 7 of the 24 cases between the results of the direct observation of teacher practice and the results of the performance evaluation. In only one of these cases were the results totally opposite: outstanding in the performance evaluation and insufficient in the result of the classroom observation. In general terms, however, the results of the observation show results equal or close to those obtained by the performance evaluation. The study is rich in details in pointing out aspects to take into account to evaluate teaching practice.

Pozos and Leyva (2019) began an analysis of the reflections of the teaching planning derived from the 2017 model of the performance evaluation. Preliminary findings from this study

> … suggest that the redesign was adequate, both in the instruments–assessment tasks and grading rubrics–and in the conditions of application, close to the daily practice of the teachers. The results allow us to understand more and better the reflective process of the teacher, compared to those of 2015. Above all, they show a more acute perception of the teachers' role in the students' learning process: they relate the results obtained by their students with their pedagogical and didactic skills and offer more precise information about their training needs. (Pozos & Leyva, 2019: 56)

Unfortunately, the analysis was interrupted due to the disappearance of the National Institute for Educational Evaluation, as described in the next section.

The entrance evaluation for teaching and that of promotion to managerial positions, in its 5 years of application, was technically consolidated and gained acceptance by the teaching profession. The instrument selected suitable applicants who were just a few years out of school and who had obtained better grades in the bachelor's degree, which showed sensitivity to attract those closest to formal education (INEE, 2015b). The foregoing, together with the findings of De Hoyos and Estrada (2018), ensures that during this time the best people entered teaching and the best were promoted to managerial positions. Although there is still no evidence to support it, there are bases to expect that the entrance and promotion evaluation had an impact, or will do so in the future, on best practices in the classroom and in school management and, as a consequence, on the learning of students. The objective procedures for entering teaching remain in the new legislation, though not in the same way as they were implemented during the 2014–2018 period.

The evaluation of teaching performance, for its part, had a bumpy start that generated great opposition among those in the teaching profession. The difficulties were due to problems of coordination between a multiplicity of actors incorporated in the institutional framework for its application and in the effective communication with the participants in terms of both its meaning and the characteristics of its implementation. They also derive from the conceptualization of a multidimensional and contextualized evaluation as the only one that could account for the multifactorial nature of teaching practice. However, as we will see later, the greatest challenge can be traced back to a history of privileges that teacher unionism had accumulated for decades in Mexico. Although the problems detected in the first teacher performance evaluation of 2015 were corrected and the evaluation model was perfected for the 2017 evaluation, and even when the evaluation was gaining acceptance by a significant proportion of the teachers, it will not be continued in our country.

9.4 The Repeal of the "Misnamed Educational Reform"

The teaching profession has always represented an important base of political support in Mexico. The teachers' union, especially the CNTE, but also the SNTE, manifested during the López Obrador campaign in 2018 its open rejection of the educational reform. López Obrador himself, as a candidate, called it "the misnamed educational reform," a phrase that spread quickly. The view of educational evaluation as punitive was the mainstay of why it was reviled, despite the fact that no teacher evaluated lost his/her job. This vision of punitive evaluation and of an educational reform hostile to teachers, which was not educational at all, but was labor-related (hence the "*misnamed* …") was appropriated by the president-elect, who, in the month of September 2018, on a tour in the state of Durango, announced its disappearance. The cry of the teachers, whenever they met at a political rally, was that of "it is going to fall, it is going to fall, the reform is going to fall." In May 2019, the already President of the Republic announced that there would no longer be evaluation of teachers and that the INEE was an imposition from abroad (Aristegui, 2019). On May 15, Teacher's Day, as a gift to the union, the reform of Article 3 of the Constitution, dedicated to education, was approved, with which the Professional Teaching Service disappeared and the INEE was replaced by a non-autonomous body, dependent on the SEP, called the National Commission for Continuous Improvement of Education.

We have already explained the complex situation of the legislation for teacher evaluation, as well as the additional actors that were incorporated into the process, such as CENEVAL and ILCE. The function of the INEE was very important but limited: It was in charge of the regulatory aspects and the corresponding supervision and evaluation of the processes related to the Professional Teaching Service. However, it was not responsible for its design or its application or the decisions that resulted from it. Despite this, the INEE was the visible face of the educational reform, and specifically of the teacher evaluation. All the responsibilities of the Professional Teaching Service were attributed to it, and the Ministry of Public Education did not consider it necessary to rectify this mistaken perception. The INEE also did not display an adequate communication in this regard for several reasons: It was difficult to explain the complexity of the distribution of functions; we did not consider it appropriate to blame others, as shared a common purpose, and a massive campaign would have been required to reach one million teachers and directors of the national education system, which was not in our budget. The result was that the communication was left in the hands of the SNTE and the CNTE: careful and quiet in the first case, noisy and combative in the second, and it was in the opposite direction to that desired: Teacher evaluation is punitive and violates teacher rights. As we have already explained, the educational reform also had the purpose of recovering the State's leadership over education, and this was interpreted as a declaration of war against the union and its dissent. The communication strategy in both unions, as already explained, was very efficient. On the other hand, this communication battle, which was not even considered as such by the INEE and probably not by the Ministry of Public Education, was clearly won by the unions.

Closely related to the above are the results of the analysis carried out by ORELAC-UNESCO in its 2016 report, which describes the tension between the two agendas of the educational reform: that of the recovery of the State's control over the education and that of teacher professionalization, or in other words, between the political and pedagogical dimension of the reform (INEE, 2017). All the technical aspects of the teacher evaluation had a clear pedagogical motivation, while the implementation of the Professional Teaching Service was definitively political. The second always prevailed over the first in impact and social perception. In 2015, there were intermediate elections of some governors, municipal presidents and deputies. In a clear violation of the autonomy of INEE, the Minister of Public Education suspended the teacher evaluation that took place during electoral times, until the election was over. Unfortunately, it was not possible to uphold the reform's historic pedagogical purpose of teacher professionalization and improvement of teaching and learning.

Far from taking teachers into account, much less defining the reform together with them, as was done in Chile, and as recommended by the literature on teacher evaluation (Martínez Rizo, 2016; Schmelkes, 2014), it was always hostile to teachers. The General Law of the Professional Teaching Service emphasizes the sanctions for non-compliance and does not refer to the substantive purposes of the evaluation. The instruments and the application of the teaching entrance evaluations were heavily guarded by the federal and state police. In 2015, the Secretariat of Public Education took over the State Institute of Public Education in Oaxaca (IEEPO)—in the hands of Section XXII of the SNTE but dominated by the CNTE—by force and appointed a general director who was never able to serve in this office because the IEEPO building was immediately taken over by the union dissidents. The aspiring teacher applicants from the states under the control of the CNTE who wanted to be evaluated were transferred clandestinely to neighboring states. The application of the performance evaluation at testing sites was heavily guarded by police in the "at risk" states—notably Michoacán, Chiapas, Oaxaca and Guerrero—and in some cases, the teachers were transferred by military helicopter to the test site to circumvent the blockades of dissident teachers that prevented access. Although no teacher was fired due to the results of their performance evaluation, around 400 teachers, who, having been summoned, did not appear for the evaluation, lost their employment. Most of these teachers have already had their positions restored since the new government took office.

The height of this hostility occurred on June 19, 2016. In the town of Asunción Nochixtlán, in Oaxaca, "hundreds of elements of the municipal, state and federal police" brutally repressed a protest against the educational reform and against the recent apprehension of two CNTE leaders. The balance was at least 8 dead and 108 injured, including a teacher (*Comisión Nacional de los Derechos Humanos* CNDH, National Human Rights Commission, 2016). With this—and this is a personal opinion of the author of this chapter—the coup de grace was given to the educational reform 2013–2019. With this event, those who maintained the punitive—now even repressive—nature of the educational reform were seemingly proven correct.

The new Constitutional Article 3 recognizes the leadership of the State over Education. It adopts a human rights approach and prioritizes the best interests of children. It recognizes teachers as fundamental agents of the educational process and establishes that "they will have the right to access a comprehensive system of education, training and updating, with feedback through diagnostic evaluations to meet the objectives and purposes of the National Educational System. It stipulates that a secondary law "shall establish the provisions of the System for the Teaching Profession and Teachers in their teaching, leadership or supervisory functions. The Federal Government will be responsible for its management and its implementation will be carried out in coordination with the federal entities…."

In some way, an evaluative procedure is maintained for entering teaching. The same secondary law establishes the procedures for "admission, promotion, and recognition of the personnel that exercise the teaching, leadership, or supervisory functions, (which) will be carried out through selection processes to which applicants attend on equal conditions as established in the law provided for in the previous paragraph, which will be public, transparent, equitable, and impartial and will consider the knowledge, skills, and experience necessary for the learning and integral development of the students. The appointments derived from these processes will only be granted in terms of said law. The provisions of this paragraph in no case affect the permanence of the teachers currently in the service."

At the same time, the *Sistema Nacional para la Mejora Continua de la Educación* (National System for the Continuous Improvement of Education, formerly the National Educational Evaluation System) is created, which will be coordinated by a decentralized public body (no longer an autonomous constitutional body) with technical, operational, budgetary, decision-making, and management autonomy, with its own legal personality and assets, not sectorized, to which it will correspond, among other things, to carry out studies, specialized investigations and diagnostic, training and comprehensive evaluations of the National Educational System. The emphasis is on the diagnostic and formative character of the evaluations.

9.4.1 Closing Words: Lessons Learned

Teacher evaluation in Mexico has gone through several stages, described in the first part of this chapter. But undoubtedly, the most intense was the one that took place between 2013 and 2019 with the educational reform of the PRI administration of Peña Nieto. The radical nature of the reform, its intensity, the strengthening, and creation of institutions to operate it and the Government's determination to regain State control by reducing the power of a union that historically had a fundamental political role in the history of the country, contributed to chart a course for teacher evaluation that was plagued with difficulties and never free of discomfort and tension. The experience was able to demonstrate the value of merit as a fundamental criterion for entering teaching and also, although to a lesser extent, for promotion to managerial positions. It is very likely that this experience will be capitalized on in the future hiring and promotion policies, making it difficult to return to the corrupt mechanisms of sale, rent, and inheritance of positions that, although they were not completely eliminated, especially in entities with the strong presence of the CNTE, were indeed considerably reduced. On the other hand, the educational reform in question had the capacity to "vaccinate" the teachers against any attempt to evaluate their performance through a system that could threaten their job security.

The reform was understood as a political decision that at the end helped the union recover spaces of control by getting the teachers to confuse the Union with the teaching profession. Although the teachers could understand the benefits of teacher professionalization, they saw themselves as victims of a threat to their job stability. This partly distorted view of the Reform was nurtured by the union itself over the years, and together with this perspective, the evaluation processes generated continuous work stress in the teaching profession. The INEE, the institution in charge of giving the process its pedagogical value, of professionalizing the teaching profession, was seen as the root cause of these years of teacher anxiety. These conditions did not allow for adequately managing the tension that the OREALC UNESCO evaluation (INEE, 2017) described between the recovery of State control and the professionalization of teachers.

A reform of these dimensions, judging by its results, cannot be carried out with such radicality and speed; it must be prepared. On the other hand, it cannot be expected to be successful without the participation and approval of those it affects: the teachers. The opportunity for the teachers to understand the advantages of their professionalization was lost; on the other hand, the face of political labor force of the teaching profession was bolstered and its nature as an educational profession was blurred.

However, the experience had the virtue of being carefully and widely documented and scrupulously evaluated. This is reported in the publications and databases found on the INEE page (https://inee.edu.mx) that is kept on the web as a historical file. Valuable lessons can be derived about novel ways of approaching the evaluation of teaching practice. There are important lessons to be learned about the creation of evaluative instruments of a diverse nature. There is also a model for evaluating teacher performance, the one from 2017, which was promising. There is information available that supports the hypothesis that teacher evaluation can effectively select the best persons to enter teaching and improve the teaching practice of those who are already within the system. There is also valuable information that, when crossed with data on academic achievement, allows the hypothesis about the relationship between good teaching and student learning to be tested.

References

Aristegui, C. (08 de enero, 2019). Andrés Manuel López Obrador: There will no longer be teacher evaluation. INEE was an imposition from abroad. Aristegui Noticias. https://aristeguinoticias.com/0805/multimedia/amlo-ya-no-habra-evaluacion-a-maestros-el-inee-fue-una-imposicion-del-extranjero-enterate/

Arnaut, A. (1993). Historia de una Profesión. Los docentes de educación primaria en México 1887–1993. El Colegio de México.

Backhoff, E., & Contreras, S. (2014). "Corrupción de la medida" e inflación de los resultados de ENLACE. *Revista Mexicana de Investigación Educativa*, *19*(63), 1267–1283. http://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S1405-66662014000400012&lng=es&tlng=es

Bracho, T., & Zorrilla, M. (2015). Perspectiva de un gran reto. En INEE. Reforma Educativa: Marco Normativo (pp. 15–38). INEE. https://www.inee.edu.mx/wp-content/uploads/2019/01/P1E101.pdf

CNDH. Comisión Nacional de Derechos Humanos. (2016). Comisión de seguimiento a los hechos ocurridos en Nochixtlán, Oaxaca, el 19 de junio de 2016. "Masacre en Nochixtlán, Oaxaca". https://www.cndh.org.mx/noticia/masacre-en-nochixtlan-oaxaca

De Hoyos, R., & Estrada, R. (01 de octubre, 2018). "¿Los docentes mejoraron? ¡Sí!, en Nexos, sección Expediente. https://www.nexos.com.mx/?p=39531

DOF. Diario Oficial de la Federación. (2013). DECRETO por el que se reforman los artículos 3o. en sus fracciones III, VII y VIII; y 73, fracción XXV, y se adiciona un párrafo tercero, un inciso d) al párrafo segundo de la fracción II y una fracción IX al artículo 3o. de la Constitución Política de los Estados Unidos Mexicanos. https://www.dof.gob.mx/nota_detalle.php?codigo=5288919&fecha=26/02/2013

de Ibarrola, M. (2018). Evaluation of teachers of basic education: Political tensions and radical oppositions. *Education Policy Analysis Archives, 26*(53). https://doi.org/10.14507/epaa.26.3819

Ducoing, P. (2019). Un acercamiento al Programa de Carrera Magisterial. En: Ducoing, P. (ed.) Programas y políticas de evaluación docente en educación básica (1993–2017) (pp. 79–141). IISUE-UNAM.

Echávarri, J., & Peraza, C. (2017). Modernizing schools in Mexico: The rise of teacher assessment and school-based management policies. *Education Policy Analysis Archives*, *25*(90). epaa.v25.2771.

Flamand, L., Arriaga, R., & Santizo, C. (2020). Reforma educativa y políticas de evaluación en México, ¿Instrumentos para abatir el rezago escolar y promover la igualdad de oportunidades? *Foro Internacional (FI), LX, núm. 2*(240), 717–753. https://doi.org/10.24201/fi.v60i2.2737

Fumagali, L., & López. N. (coords.) (2015). Evaluación internacional de los procesos de evaluación de ingreso y promoción al Servicio Profesional Docente en Educación Básica y Educación Media Superior en México, 2014–2015. INEE. https://www.inee.edu.mx/portalweb/suplemento12/evaluacion-ingreso-y-promocion-al-spd-2014-2015.pdf

Gluyas, R., & González, Z. (2014, September). Universal evaluation: An invitation to the creation of innovative models for teacher training. *Journal of Case Studies in Education, 6*. https://www.aabri.com/manuscripts/131755.pdf

Guzmán Marín, F. (2018). La experiencia de la evaluación docente en México: Análisis crítico de la imposición del servicio profesional docente. *Revista Iberoamericana de Evaluación Educativa, 11*(1), 135–158. Recuperado de https://doi.org/10.15366/riee2018.11.1.008

INEE. Instituto Nacional para la Evaluación de la Educación. (2015a). Reforma Educativa. Marco Normativo. México: INEE. https://www.inee.edu.mx/wp-content/uploads/2019/01/P1E101.pdf

INEE. (2015b). Los Docentes en México. Informe 2015b. INEE. https://www.inee.edu.mx/publicaciones/los-docentes-en-mexico-informe-2015b/

INEE. (2017). Evaluación del desempeño de docentes, directivos y supervisores en educación básica y media superior de México. Análisis y evaluación de su implementación 2015–2016. Informe final. México: OREALC/UNESCO. https://www.inee.edu.mx/wp-content/uploads/2019/01/P1F209.pdf

INEE. (2018a). La Educación Obligatoria en México. Informe 2018a. INEE. https://www.inee.edu.mx/wp-content/uploads/2018a/12/P1I243.pdf

INEE. (2018b). Encuesta de satisfacción de los procesos de evaluación de ingreso y promoción en educación básica y media superior 2017. INEE. https://www.inee.edu.mx/wp-content/uploads/2018b/12/P1F224.pdf

INEE. (2018c). Resultados de la encuesta de satisfacción. Evaluación de Desempeño. INEE. https://www.inee.edu.mx/wp-content/uploads/2018c/12/P1F225.pdf

INEE. (2019a). Concurso de Oposición para el Ingreso al Servicio Profesional Docente en Educación Básica y Media Superior para el ciclo escolar 2017–2018. Informe de supervisión y de observación. INEE. https://www.inee.edu.mx/wp-content/uploads/2019a/04/P1F229.pdf

INEE. (2019b). La Educación Obligatoria en México. Informe 2019b. https://www.inee.edu.mx/wp-content/uploads/2019b/04/P1I245.pdf

INEGI. (2020). Percepción de Inseguridad Pública. https://www.inegi.org.mx/temas/percepcion/

LGSPD. (2013). Ley General del Servicio Profesional Docente. https://www.dof.gob.mx/nota_detalle.php?codigo=5313843&fecha=11/09/2013

Martínez Rizo, F., & Blanco, E. (2010). La evaluación educativa: experiencias, avances y desafíos. En: Alberto Arnaut & Silvia Giorguli (coords). Los grandes problemas de México VII. Educación. El Colegio de México.

Martínez Rizo, F. (2016). La evaluación de docentes de educación básica. Una revisión de la experiencia internacional. INEE. https://www.inee.edu.mx/wp-content/uploads/2018/12/P1C233.pdf

Plá, S. (2019). Calidad educativa: historia de una política para la desigualdad México: IISUE-UNAM.

Pozos, P., & Leyva, Y. (2019). La reflexión como eje fundamental del portafolio de evaluación de la práctica docente. *INEE Red 5*(13), mayo-agosto, 50–69.

Santibáñez, L., et al. (2006). Haciendo camino: análisis del sistema de evaluación y del impacto del programa de estímulos docentes Carrera Magisterial en México. SEP- RAND Corporation.

Schmelkes, S. (2014). La Evaluación del Desempeño Docente: Estado de la Cuestión. En: OREALC/UNESCO Santiago. *Temas Críticos para Formular Nuevas Políticas Docentes en América Latina y el Caribe: El Debate Actual* (pp. 154–185). UNESCO, Oficina Regional para América Latina y el Caribe.

Weiss, E., Block, D., Civera, A., Dávalos, A., & Naranjo, G. (2019). La Enseñanza en Educación Básica: Análisis de la práctica docente en contextos escolares. INEE. https://www.inee.edu.mx/wp-content/uploads/2020/02/P1F233.pdf

Chapter 10
Teacher Evaluation in Peru: Prospects and Challenges

Giuliana Espinosa and Liliana Miranda

Abstract Since 2000, with the recognition that teaching in Peru—which was greatly devalued at the time—needed to be professionalized, teacher evaluation policies have been initiated within a broader context of concern for teacher development. However, it was not until the implementation of the *Ley de Reforma Magisterial* (LRM, Teacher Reform Law) of 2012 that teacher evaluation policies were clearly and more systematically implemented. The LRM unified all the teachers in the public system under the *Carrera Pública Magisterial* (CPM, Public-School Teaching Career) regime governed by the principle of merit. Under this framework, a complex system of teacher evaluations was established, which regulates, among other processes, the admission, permanence and promotion of teachers within the career regime. The advances and achievements of the new evaluation system are significant, and their impacts are unquestionable; however, the implementation process has revealed a series of challenges for teacher development policy that must be addressed to promote the desired professionalization of teachers and rethink the role of evaluation within it. This chapter analyzes, first, the context of creation and development of the teacher evaluation system in Peru. Then, the evaluation system developed in the framework of the CPM is presented, with special attention to its purposes, instruments, characteristics and results. Finally, a full accounting of the system is taken along with some of the challenges that this new phase is confronting from a broader policy vision of teacher development.

G. Espinosa (✉)
Innova Teaching School—ITS, Lima, Peru
e-mail: gespinosa@its.edu.pe

L. Miranda
Grupo de Análisis para el Desarrollo—GRADE, Lima, Peru
e-mail: lmiranda@grade.org.pe

J. Manzi et al. (eds.), *Teacher Evaluation Around the World*,
Teacher Education, Learning Innovation and Accountability,
https://doi.org/10.1007/978-3-031-13639-9_10

10.1 Background of the Evaluation System

Teacher evaluation represents one of the most significant efforts of the Peruvian educational system in recent years although it is still in a consolidation phase. Originally, this public action was linked to a broader State policy to rethink teacher development in an effort to recover the social prestige of being a teacher and provide the career with professionalization tools to attract and retain motivated, committed and competent teachers. This stemmed from the commitment to strengthen public basic education for the benefit of the more than six million students served by it.

To provide background about the context in which the processes mentioned are situated, some characteristics of the Peruvian teaching system are briefly presented in what follows, taking into account that the current CPM regime includes teachers who work in the public sector in all the modalities of Basic Education and in the Technical Productive programs.

In Peru, there are almost 540,000 teachers who serve Basic Education students[1] in both public and private institutions (see Table 10.1). Most of the teachers work in the Regular Basic Education modality (95%), in institutions managed by the State (72.6%) and in institutions located in urban areas (69.3%).

The population of teachers in the public sector is almost two-thirds women (62.8%), with an average age of 46, with 93.6% holding a teacher's degree or a degree in education. The employment status of *civil service teacher* (permanent teacher) is held by 58% of the teachers, while 15.4% report that they have another occupation in addition to teaching (Minedu, 2021).

In Peru, the revaluation of the teaching career was the product of a long process of reflection, analysis and political negotiation that began in the context of the reestablishment of democratic order, after the fall of the government of Alberto Fujimori, at the beginning of the new century. The new democratic scenario demanded citizen participation in the design of public policies, thus opening spaces for social dialogue that later converged in the 2002 *Acuerdo Nacional* (AN, National Agreement). This pact brought together the most representative political parties and civil society organizations and established a set of 35 State Policies, one of which is related to education. It proposed the strengthening and reassessment of the teaching career through a social pact that ensures optimal professional training and greater resources for this purpose.

While the AN was being developed, between 2001 and 2002, the Ministry of Education (Minedu) commissioned a set of studies on various aspects related to the profession of public teaching and teacher professional development. These studies recognized the need to introduce a teaching career regime based on merit. Thus, in 2003, the *Ley General de Educación* (LGE, General Education Law) was enacted, incorporating these conclusions. This Law puts the student at the center of all actions

[1] Made up of the modalities of Regular Basic Education, Alternative Basic Education, Special Basic Education. In addition, teachers of Technical Productive Education are usually considered in this group, which is basic occupational education that is not recognized within the higher education system.

Table 10.1 Peru: Number of teachers[2] by type of management and area, according to modality and educational level, 2020

Modality and educational level	Total	Management of the educational service		Geographical area of educational service	
		Public	Private	Urban	Rural
Regular basic	513,816	373,103	140,713	349,578	164,238
Initial[a]	95,306	63,922	31,384	67,291	28,015
Primary	213,618	154,199	59,419	142,242	71,376
Secondary	204,892	154,982	49,910	140,045	64,847
Alternative basic	12,584	8738	3846	11,994	590
Special basic	4470	4138	332	4283	187
Technical productive	9036	5776	3260	8224	812
Total	539,906	391,755	148,151	374,079	165,827
%	100.0	72.6	27.4	69.3	30.7

Adapted from Minedu (2020)

[a] Excludes community education promoters in charge of out-of-school programs

within the educational system and establishes for the first time that the Public-School Teaching Career must be governed by the principle of merit, thus making entry to the system, permanence and promotion within it subject to evaluation processes. Until then, civil service teachers in the public sector were governed by the *Ley del Profesorado* (LP, Teachers' Law)[3] of 1984, with a five-level scale with little remuneration difference between the scale's steps. Promotions occurred with fulfillment of service time, and there was almost absolute job stability (see Fig. 10.1).

In 2007, the *Ley de Carrera Pública Magisterial* (LCPM*, Public Teaching Career Law*)[4] was passed that established what was specifically indicated in the *Ley General de Educación* (LGE, General Education Law) and the new work regime for public school teachers based on merit. This new career regime meant more attractive salaries with greater salary differences in each of its five teaching levels. At the same time, it required teachers to be evaluated with consequences that could even lead to dismissal from the teaching profession as a result of poor performance. The transition from teachers who were in the old regime to the new one was voluntary. In this way, the coexistence of the two work regimes for teachers was established.

Approval of the Public Teaching Career Law occurred amid fierce disputes by the political forces present in the Parliament and a national strike by the teachers' union. The legitimacy of the central norm that governs the professional life of teachers was affected from its origin and the implementation process was no less complex. In addition, various questions were raised about the reliability and validity of the

[2] Corresponds to the sum of the number of people who perform teaching, managerial or classroom work, in each educational institution, without differentiating between full and part-time work.

[3] Law No. 24029 (1984).

[4] Law No. 29062 (2007).

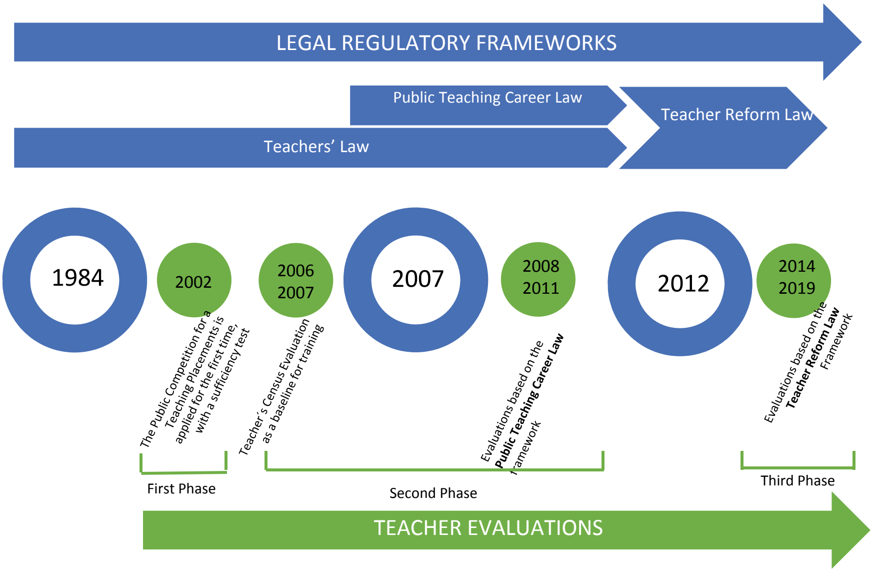

LEGAL REGULATORY FRAMEWORKS
Public Teaching Career Law
Teacher Reform Law
Teachers' Law
1984
2002
2006
2007
2007
2008
2011
2012
2014
2019
The Public Competition for a Teaching Placements is applied for the first time, with a sufficiency test
Teacher´s Census Evaluation as a baseline for training
Evaluations based on the **Public Teaching Career Law** Framework
Evaluations based on the **Teacher Reform Law** Framework
First Phase
Second Phase
Third Phase
TEACHER EVALUATIONS

Fig. 10.1 Legal framework and phases of teacher evaluations. *Source* Own elaboration based on Cuenca (2020)

evaluations, the administration of the teaching profession, as well as the negative effects on the work environment of educational institutions as teachers who did the same work on the same schedule received different remuneration (Herrero, 2012).

As of 2012, only 20% of the public sector's civil service teaching staff were in the new career regime and the evaluations to establish teacher permanence had not yet been implemented (Cuenca, 2012). The designation to other positions was carried out through innumerable normative devices. For these reasons, in November 2012 under a new government, the Congress approved the Teacher Reform Law (LRM) with the purpose of "establishing a single labor regime for public sector teachers and developing a career path based on teaching merit" (CNE, 2017, p. 106).

In addition, the Teacher Reform Law organized the Teaching Career regime in eight scales and four areas of work performance so that teachers had access to a career path with opportunities for improvements in remuneration and in their working conditions, along with diverse possibilities for professional development. This Law established specific requirements and minimum periods required for each one of the teaching scales. The areas of job development that it recognized included (i) pedagogical management, (ii) institutional management, (iii) teacher training and (iv) innovation and research.

Finally, the LRM explicitly stated that all evaluations should have a fundamentally educational purpose and should support the Minedu and regional bodies by identifying the training actions relevant to promote continuous teacher improvement within the framework of the different areas of job performance within the teaching profession. However, this formative intentionality was set against the summative orientation of the evaluation—closer to the approach of *accountability*—which impacted on significant aspects of teacher development such as career advancement and therefore on teachers' remuneration, or, even more, on potential job loss. Reconciling both purposes constituted one of the challenges of the teacher evaluation policy in Peru.

10.1.1 Phases of Teacher Evaluations

According to Cuenca (2020), teacher evaluations in Peru can be classified into three phases. These phases respond to the regulatory frameworks that have driven the development of the system from the beginning of the century to the present (see Fig. 10.1).

The first phase included the evaluation carried out in 2002, within the framework of the Teachers' Law, in which the Ministry of Education (Minedu) included, for the first time, meritocratic criteria in the *Concurso Público para el Nombramiento de Plazas Docentes* (Public Competition for the Credential for Teaching Placements). The proposal added to the traditional evaluation criteria, the academic qualification through a Professional Sufficiency Test[5] and a personal interview (Lynch, 2006; Piscoya, 2005). This evaluation occurred within a scenario of rejection and resistance by the union.

The second phase of evaluations took place in the period from 2006 to 2011. Its starting point was the Census Evaluation[6] for basic education teachers to obtain information for the design of training actions. This decision generated criticism from specialists due to its improvisation and opposition from teachers, which is why coverage of the evaluation reached only 66% of the teachers who had been planned to evaluate (Secretariat of Strategic Planning, 2007). In this same period, within the framework of the implementation of the Public Teaching Career Law, eight evaluations were carried out for admission to the public teaching profession, both for the credentialing of new teachers and for the incorporation of in-service teachers to the profession (Cuenca, 2020). According to the National Education Council (CNE), the evaluation instruments used generated serious "questions about their quality and relevance in a context of tension and conflict with Peruvian teachers" (CNE, 2019; p. 60).

The third and last phase of evaluations began in 2014, continues to date and was developed within the framework of implementing the Teacher Reform Law. This

[5] The *Prueba de Suficiencia Profesional* was national and diversified according to the levels, specialties and modalities of the placement in the Competition. It consisted of multiple-choice questions grouped into three areas: (i) general culture, (ii) pedagogical and (iii) teaching aptitude.

[6] The *Evaluación Censal* was carried out in 2006 and 2007 and comprised two parts, both multiple choice. The first consisted of three subtests of (i) reading comprehension, (ii) mathematics and (iii) general knowledge of the curriculum. The second part included knowledge of the curriculum for each educational level.

standard provides for a set of regular evaluations that are detailed in the second section of this chapter. It should be noted that prior to the implementation of the regular evaluations a set of exceptional evaluations were developed with the objective of ordering and facilitating the transition to the new regime.

Unlike the previous evaluation, in this third phase, the evaluations were developed with transparency and appropriateness,[7] as well as with a high level of teacher participation and little conflict, with the exception of the first Teacher Performance Evaluation done in 2017[8] (Cuenca, 2020; Cuenca & Vargas Castro, 2018).

10.1.2 Benchmarks or Evaluation Standards

From the perspective of strengthening the professional competencies of the Peruvian teaching profession, the Teacher Reform Law established that the design of teacher training and evaluation processes must have as a reference the *Marco de Buen Desempeño Docente* (MBDD, Good Teaching Performance Framework). This educational policy instrument proposed a vision of teaching and defined the professional competencies required of all Basic Education teachers (Minedu, 2013).

The construction of the Good Teaching Performance Framework in Peru was a technical effort, but above all, it was participatory. As Cruz-Aguayo et al. (2020) highlighted, the relevance of educational system tools that make explicit what a teacher should know and should know how to do is based not only on technical dimensions, but also on the legitimacy that it gives to the development of teaching policy.

The Good Teaching Performance Framework initiative came from civil society. It was created during the period 2009 and 2012 under the leadership of the National Education Council and Educational Forum[9] through the Inter-institutional roundtable for Good Teaching Performance. This space brought together a broad group of State and civil society institutions, including the teachers' union and the teachers' association. From there, it was proposed to contribute to the discussion and construction of a consensus on the characteristics of those who teach in basic education in the various contexts of the country. To do this, they designed a strategy that included conducting

[7] The only setback occurred in the first competition for access to vacancies in managerial positions in 2013—designed to be applied by computer—which was suspended twice due to system failures. After this problem, Minedu decided to apply the test in pencil and paper format.

[8] The breach of the electoral promise made by President Pedro Pablo Kuczynski (elected in July 2016) of a significant increase in teacher salaries was the trigger for a complex union conflict between regional leaders opposed to the *Sindicato Unitario de Trabajadores de la Educación del Perú* (Sutep, Unitary Union of Education Workers of the Peru), which led to different strikes. The regional leaders demanded, among other measures, the suspension of the Teacher Performance Evaluation due to its consequences for the job security of teachers. Finally, Minedu proceeded to intensify its communication campaign regarding the characteristics of the performance evaluation, made some adjustments to it and kept the evaluation standing.

[9] Civil Society Association.

studies, workshops and consultations with teachers and regional and national meetings. At the end of 2011, the Roundtable presented to Minedu a proposal for a framework for guiding good teaching performance. This approach was reviewed by the Ministry, through the convocation of a panel of experts and different civil society organizations—including the teachers union—and finally, in December 2012, the guidelines called "Framework of Good Teaching Performance for Regular Basic Education Teachers" were approved[10] (Minedu, 2013).

The participatory character of the framework's creation assured that the instrument is known and valued positively by a majority of Peruvian teachers. Indeed, according to information from the National Survey of Teachers for 2014, "at least 77% of Peruvian teachers, public and private, knew about the Framework (…) and the vast majority recognized its usefulness to improve classroom work (89%), as inputs for teacher training and evaluation (86% and 80%), and even as an instrument for social revaluation of the profession (83%)" (Cuenca, 2020, pp. 14–15).

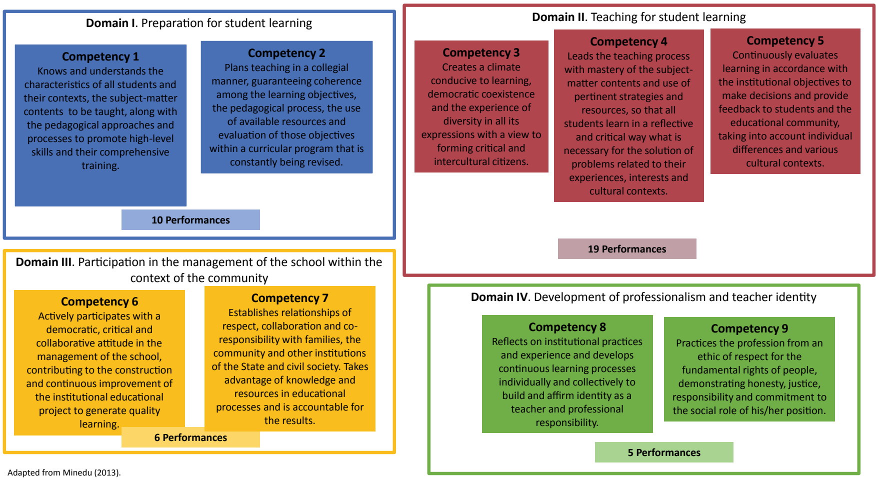
As for the conceptual approach used in the Good Teaching Performance Framework (MBDD), this was inspired, among other approaches, by Chile's Framework for Good Teaching, the basis of which was the work carried out by Danielson in the United States (Vázquez et al., 2014). Guerrero (2011) points out that this was developed by collecting various theoretical approaches that were based on four premises: (i) teaching as a relational profession where the bond is essential, (ii) the teacher as a capable professional to discern and make decisions; (iii) professional performance as practice and action and (iv) the teaching functions assigned by the Law and the National Educational Project in force as a framework.

The MBDD contemplates a hierarchical structure of three categories: four domains, nine competencies and forty performances (see Fig. 10.2). A domain is defined as the field of teaching that groups together a set of professional performances that favorably affect student learning. Competencies are understood as the ability to solve problems and achieve objectives, not only as the ability to put knowledge into practice. Finally, the MBDD considers performances to be observable actions that can be described and evaluated and that demonstrate competence (Minedu, 2013).

Although the MBDD presents a general profile on the professional competencies required of all teachers, it was necessary to specify them to enhance their use as a common basis for teacher development policies, especially for the design of a new in-service training model, and to serve the need to have relatively stable references that account for a progression in the development of these competencies. This progressive approach would make it possible to mark milestones in the training process of competencies, to focus on the areas of proximal development for teachers and, at the same time, establish differentiated levels of performance in the face of evaluation processes (*Dirección de Formación Docente en Servicio*, Directorate of In-Service Teacher Training, 2017).

Within the described framework, Minedu approved the MBDD Standards in Progression of Professional Competencies in 2020 with the purpose of supporting

[10] Ministerial Resolution No. 0547-2012-ED.


Domain I. Preparation for student learning
Competency 1
Knows and understands the characteristics of all students and their contexts, the subject-matter contents to be taught, along with the pedagogical approaches and processes to promote high-level skills and their comprehensive training.
Competency 2
Plans teaching in a collegial manner, guaranteeing coherence among the learning objectives, the pedagogical process, the use of available resources and evaluation of those objectives within a curricular program that is constantly being revised.
10 Performances
Domain II. Teaching for student learning
Competency 3
Creates a climate conducive to learning, democratic coexistence and the experience of diversity in all its expressions with a view to forming critical and intercultural citizens.
Competency 4
Leads the teaching process with mastery of the subject-matter contents and use of pertinent strategies and resources, so that all students learn in a reflective and critical way what is necessary for the solution of problems related to their experiences, interests and cultural contexts.
Competency 5
Continuously evaluates learning in accordance with the institutional objectives to make decisions and provide feedback to students and the educational community, taking into account individual differences and various cultural contexts.
19 Performances
Domain III. Participation in the management of the school within the context of the community
Competency 6
Actively participates with a democratic, critical and collaborative attitude in the management of the school, contributing to the construction and continuous improvement of the institutional educational project to generate quality learning.
Competency 7
Establishes relationships of respect, collaboration and co-responsibility with families, the community and other institutions of the State and civil society. Takes advantage of knowledge and resources in educational processes and is accountable for the results.
6 Performances
Domain IV. Development of professionalism and teacher identity
Competency 8
Reflects on institutional practices and experience and develops continuous learning processes individually and collectively to build and affirm identity as a teacher and professional responsibility.
Competency 9
Practices the profession from an ethic of respect for the fundamental rights of people, demonstrating honesty, justice, responsibility and commitment to the social role of his/her position.
5 Performances

Adapted from Minedu (2013).

Fig. 10.2 Domains of the good teaching performance framework

individual or collective reflection of teachers on their own practice within a framework of lifelong learning and at levels of increasing complexity. Likewise, it sought to contribute to the coordination and synergy of the teacher training system with the other components of teacher development policy.[11]

Thus, the Standards establish three levels of progression in the competencies of the MBDD. They start from a hypothesis based on evidence of how these competencies are typically developed in practicing teachers. Table 10.2 presents an example of progressive levels in a competency.

Since their approval, the Standards have been used in the design of the new initial training curricula approved between 2019 and 2020. However, it is still too early to make a more comprehensive assessment of the use that Peruvian teachers are making of these as a guide to reflect and evaluate their own teaching practices. To what extent it is guiding policies to strengthen the teaching profession is also an open question. Likewise, evaluations designed after its publication have not yet been implemented.

10.2 Characteristics and Results of the Peruvian Teacher Evaluation System

The teacher evaluation system constitutes a fundamental pillar of the teacher development policy covered by the Teacher Reform Law, which proposes *merit* as one of the basic principles of the new teacher labor regime: "The admission, permanence, remuneration improvements and promotions in public school teaching are based on the merit and ability of teachers" (Minedu, 2018; p. 17).

[11] Vice-Ministerial Resolution No. 005-2020.

Table 10.2 Progressions of competency 3 of the good teaching performance framework

Domain 2. Teaches for the learning of all students		
Competency 3. Creates a climate conducive to learning, democratic coexistence and the experience of diversity in all its expressions with a view to forming critical and intercultural citizens		
Capacities. (i) Generates an environment of respect, trust and empathy based on the valuation of diversity; (ii) promotes the involvement of all students in the learning process and in the classroom in general; (iii) regulates coexistence based on the concerted construction of standards and the democratic resolution of conflicts		
Progression		
Level I	**Level II**	**Level III**
Creates a climate characterized by respectful relationships and empathy with and between the students, taking into account their characteristics and intervening in cases of discrimination that occur in the classroom Promotes the involvement of all students in the learning process, motivating them to participate, supporting their opinions about issues related to the classroom environment and expressing themselves, confident in their possibilities to learn. Consistent with this, directs the process of defining standards of coexistence oriented to promote the common good and regulate coexistence based on these. When conflicts arise in the classroom, brings the implicated parties together and proposes solutions to them	Creates a climate characterized by respectful relationships and empathy with and between the students, promoting relationships based on trust, in which ties of solidarity and cooperation among group members are created. To achieve this, encourages recognition and expression of the different identities in the classroom and intervenes and manages situations of discrimination Promotes the involvement of all students in the learning process, motivating them to participate, starting from their particular characteristics, proposing challenging tasks and giving them opportunities to intervene in decision making in matters related to classroom life. Consistent with this, builds with them standards of coexistence relevant to the reality of the group, oriented to promote the common good. Regulates classroom coexistence based on these standards, ensuring that students understand why it is important to comply with the established agreements and what the consequences are of transgressing them When conflicts arise, promotes active participation of the parties involved to seek democratic solutions	Creates a climate characterized by respectful and empathetic relationships with and among students, promoting their responsibility in forming a community established on bonds of solidarity and cooperation, which automatically acts in cases of discrimination. Promotes the expression of different identities and reflection on the implications of a coexistence in diversity. Promotes the involvement of all students in the learning process, motivating them to participate based on their particular characteristics, proposing challenging tasks and giving them opportunities to intervene in decision making on matters related to classroom life. Uses various mechanisms to balance the level of student participation. In addition, builds with them rules of coexistence relevant to the reality of the group, oriented to the common good. Works to mediate the management of coexistence in the classroom to orient students and encourage them to be active in managing their interpersonal relationships. When conflicts arise, employs strategies that are pertinent to the nature or complexity of conflicts and promotes the active participation of the group in their resolution, regardless of those who have been involved in the conflict

Own elaboration based on Vice-Ministerial Resolution No. 005-2020 that approves the Standards

The Law establishes the path within the Public-School Teaching Career in which evaluations play a fundamental role and are connected with incentive systems and formative activities for teacher development. In this way, the career path begins with entry at the first level of the career scale (of the 8 of which it is composed) and the position as classroom teacher (starting position)[12] based on a public competition. Upon entering, the teacher goes through an induction process defined by the Minedu. This induction process is not part of the evaluation period but is proposed as a space for getting accustomed to the public educational system, during which newly appointed teachers are accompanied by a more experienced teacher, who guides them in their adaptation process.

After the induction stage, each teacher is mandatorily evaluated on his/her performance periodically (every five years) in order to remain in the career. In case of disapproval, access to a strengthening program is provided and the teacher has up to two more opportunities for evaluation. If the teacher passes the performance evaluation and remains in the career, he/she accumulates time of service to take part in competitions for promotion. The promotion represents an improvement in remuneration and enables teachers to access higher-ranking positions within the Public-School Teaching Career, also through competition. Only failure on three consecutive performance evaluations leads to dismissal from the career.

Thus, the Teacher Reform Law recognizes four types of evaluations, whose administrative functions determine admission, permanence (evaluation based on teacher performance), upward movement on the career scale and access to other positions. Next, we describe the system evaluations in which the teachers who teach in the classroom participate: admission, promotion and performance.[13]

10.2.1 Evaluation for Admission

The admission evaluation is the basis for obtaining a civil service position into the system and is used to select the new teachers for the Public-School Teaching Career. In Peru, teaching positions are filled with civil service teachers (permanent teachers) or with contracted teachers (temporary teachers). The latter do not enjoy the same benefits as the former and only have a contractual relationship for a specified period, usually one year. In this way, the civil service positions are quite desirable because of their stability and working conditions. This high demand, added to the need to attract better teachers to the career, has made this evaluation the most demanding and complete in the system.

In the design of this evaluation, it is assumed that the system for initial teacher training is deficient and does not ensure basic skills in its graduates. Therefore, it not

[12] This position belongs to the area of pedagogical management.

[13] Given the nature of this book, admission and performance evaluations have not been included for other positions on the Career scale such as managerial positions of educational institutions or specialists and managers of decentralized educational management bodies.

only assesses specialized professional knowledge and skills but also basic skills. In this way, the model comprises six instruments, which are applied in two stages, the first at the national level through the so-called *Prueba Única Nacional* (PUN, Single National Test), which is classificatory, and the second at a decentralized level and developed at those schools that have an open position.

This national entry level test has three parts that measure (i) reading comprehension, (ii) logical reasoning and (iii) knowledge of the specialty (curricular, disciplinary and pedagogical). Although this last part is the one with the greatest weight in the score, the applicant must pass the cut-off points of each part to move on to the next stage (Minedu, 2019). These cut-off points are established as percentages of correct answers with respect to the total number of items. The difficulty of the test is regulated according to the judgment of experts who created the matrix using as a reference the basic education curricular framework and the Good Teaching Performance Framework. In areas that have already been evaluated, the team that develops the tests also considers the psychometric behavior of items and instruments applied in previous evaluations.

The score on this admission test is obtained from the simple sum of the scores of its component parts, calculated as presented in Table 10.3.

Applicants who qualify choose the places of their interest on a platform and are assigned to them based on merit according to the score obtained on the admission test. In the decentralized stage which is the responsibility of evaluation committees chaired by school administrators, three instruments are applied: (i) a checklist of the qualifications for a professional standardized trajectory, (ii) a semi-structured interview designed by the committee itself to evaluate affinity to the project of the educational institution and (iii) an observation of classroom performance. Of these instruments, the one with the greatest weight in the score is classroom observation. It should be noted that the instruments are applied following the manuals and guidelines that Minedu provides to the members of the Committee. In addition, it is requested that the same members are in charge of applying the same instrument to all applicants for a position so that applicants are evaluated by the same judges. Finally, the winner

Table 10.3 Composition of the single national test

Construct evaluated	Number of questions	Value of each question	Maximum score	Cut-off score to move to decentralized stage
Reading comprehension	25	2	50	30
Logical reasoning	25	2	50	30
Specific pedagogical knowledge of the specialty	40	2.5	100	60

Adapted from Minedu (2019)

Table 10.4 Credentialing competition results, 2015–2019[14]

Public school teaching career admission competition	Positions available	Candidates evaluated	Classified PUN	Teachers who entered (covered places)	Classified (%)	Entered (%)	Positions covered (%)
2015	19,631	192,397	25,109	8137	13.1	4.2	41.4
2017	37,201	208,026	22,115	10,932	10.6	5.3	29.4
2018	35,915	194,556	24,044	10,120	12.4	5.2	28.2
2019	24,590	212,456	15,874	4554	7.5	2.1	18.5
Average percentage of credentialing processes					10.9	4.2	23.4

Source Public competitions for entrance into CPM, 2015–2019. Teacher Evaluation
Author created

of each placement is determined by adding the scores from both stages to establish each candidate's order of merit (Minedu, 2019).

To date, four competitions to credential teachers have been carried out. These competitions were originally planned to be held every two years; this provision was modified to speed up the credentialing of teachers so that as of 2017 these competitions have been held annually, as shown in Table 10.4.

As can be seen, the admission evaluation operations are massive processes in which around 200,000 people participate annually. The number of applicants far exceeds the number of places available in each competition. However, most of the positions remain unfilled. Moreover, there is a worrying trend that the percentage of places covered is decreasing. For example in the 2019 evaluation, only 1 out of every 5 places was filled. This phenomenon is partly explained by the low rates of passing the exam at the national qualifying stage, which fluctuate between 7 and 14%. Hence, this first filter is demanding for the population of teachers evaluated and implies the declassification of the majority of participants.

But there is also an effect related to the selection of places. According to Bertoni et al. (2020): "The vacancies without applicants are located in the most disadvantaged areas of the country, as well as in rural areas and in areas with a lower socioeconomic level" (p. 31). Indeed, a significant percentage of rural and remote places are not chosen by any applicant, while the few applicants who pass the qualifying stage tend to concentrate in the positions located in urban areas and in the main cities of the country, effectively eliminating each other due to the nature of the competition (Table 10.5).

Although there are incentives proposed from the Teacher Reform Law for teachers to choose rural placements, such as a monetary bonus and the possibility of ascending more quickly with the reduction of time of permanence required on each level of the teaching scale, the more rural the placement, the less likely it is to be selected.

[14] Because of the national health emergency situation caused by the Coronavirus pandemic, it was not possible to perform the evaluation for the year 2020.

Table 10.5 Percentage of vacancies in credentialing competitions by area, 2015–2019

Environment	Positions	2015 (%)	2017 (%)	2018 (%)	2019 (%)
Urban	Selected	87	79	75	75
	Covered	72	44	41	28
Nearby rural	Selected	69	68	70	66
	Covered	56	35	35	24
Distant rural	Selected	35	44	51	39
	Covered	25	21	20	11
Total	Selected	53	57	61	54
	Covered	41	29	28	19

Source Public Competitions for Entrance to the CPM, 2015–2019. Teacher Evaluation
Author created

There is also an element of no less complexity that especially affects the placements located in schools classified as Intercultural and Bilingual Education (IBE), since for these places an additional requirement is that the applicant has mastery of the native language of the area. This has led up to 70% of IBE vacancies remaining open. Thus, the system fails to attract enough candidates for the most remote, rural and bilingual areas, which perpetuates the gaps within the public school as these schools end up being served by hired teachers who have not passed the evaluation process (Bertoni et al., 2020).

In summary, by 2020 of the little more than 234,000 teachers who are in the Public-School Teaching Career program only 14% have entered under the demanding standards of this new system. In turn, these 234,000 represent 58% of the 402,000 teachers who work in the public system[15]; the rest are hired on a temporary basis and are not part of the Public-School Teaching Career program. After more than five years of implementation of the Teacher Reform Law for credential evaluations, these low percentages have begun to generate some discussion about the level of demand of the system. This has led to demands from the teachers union to review the design of the admission evaluation, focusing on the requirements of the admission exam, which few can pass and on the subsequent processes of placement selection.

For its part, the National Educational Project to 2036 prepared by the National Education Council indicated that the introduction of professional merit as a fundamental criterion for admission has been a very important step. However, given the starting situation, it has led to a change in the composition of the teaching body according to its contractual regime, which has led to a growing proportion of teachers with temporary contracts. This problem is not solved by lowering the barriers to entry or holding more competitions, but by proposing a comprehensive policy that

[15] Data obtained from the Nexus System of Administration and Control of Teacher Placements, April 2021.

addresses the difficulties in an integrated and coherent manner (National Education Council, 2020).

10.2.2 Evaluation for Moving up the Scale

During 2014 and 2015, using exceptional evaluations of relocation on the teaching scale, the Minedu was able to locate all the teachers on a scale corresponding to the new regime set by the Teacher Reform Law (LRM). Once relocated, the period of regular annual promotion evaluations began, in which moving up one step at a time on the scale is allowed after having met the minimum service time established in the previous scale step (see Fig. 10.3) and the other requirements established by the same law (LRM).

The law declares that the purpose of this evaluation is to promote the social and professional recognition of teachers, to grant them salary improvements and to identify their training needs. Like the other Public-School Teaching Career program competitions, moving up the scale requires two stages. A national test that assesses specialist knowledge and the decentralized stage where the career path is reviewed with a standardized qualification checklist, as described below.

Given a budget approved by the Ministry of Economy, the number of vacancies that will be put up for competition is determined, which are distributed in a relatively proportional way to the number of teachers qualified to be promoted by scale, modality and region. While in the case of the higher scales the competence is national, in the lower scales, the competition is limited to teachers from the same region. It should be noted that the test that teachers take varies according to specialty and not according to scale, since what determines the difference between scales is the cut-off point required to pass the test and classify: The higher the scale, the higher is the minimum correct rate required. Similar to the credentialing process, these cut-off

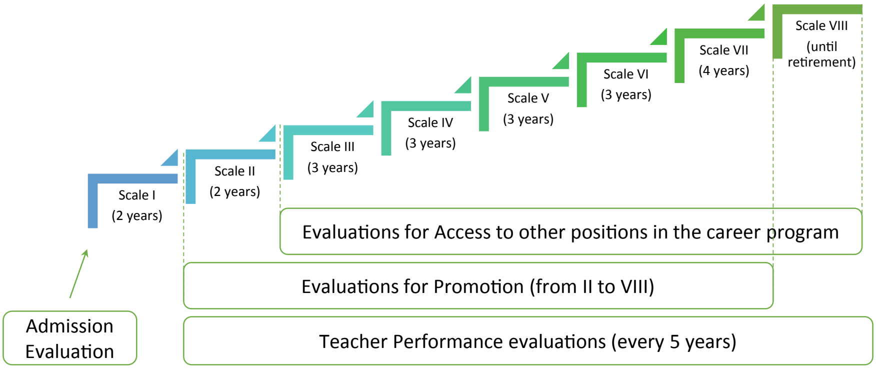


Fig. 10.3 Public-school teaching career and evaluations of the teacher reform law. Author created

Table 10.6 Results of the competition for promotion, 2016–2019

Promotion competition	Promotion vacancies	Evaluated	Classified/placed	Promotion vacancies covered by a winner	Classified/placed (%)	Positions filled (%)
Promotions 2016	40,000	50,484	8979	8506	18	21
Promotions 2017	28,320	123,490	49,648	27,963	40	99
Promotions 2018	26,307	132,425	25,341	22,846	19	87
Promotions 2019	22,243	98,450	21,628	21,101	22	95

Source Public Competitions for Scale Promotions 2016–2019. Teacher Evaluation
Author created

points are defined by the team from the Ministry of Education as a function of the criteria of judges.

In the decentralized stage, the evaluation committees are constituted by authorities and specialists of the system at provincial levels. They verify the teaching qualifications based on the documents that certify the fulfillment of the requirements and provide evidence of the training, experience and merits of the teachers who compete. Table 10.6 shows the results of the four promotion exams completed to date.

The previous results show that, with the exception of 2016, the promotion competitions managed to fill the vast majority of promotion vacancies offered. As for 2016, it should be noted that during that year only vacancies were submitted to the competition to move from the first to the second scale level because, at that time, only teachers on the first level of the scale fulfilled the time of service that enabled them to compete. For this reason, approximately half the number of teachers appeared for this evaluation than for the following competitions in which more scale levels were enabled. On the other hand, the results of the first evaluation made it possible to regulate the difficulty of the test in the following competitions to facilitate more people passing the qualifying stage and filling the available vacancies.

By 2021, more than 80,000 teachers from the system have managed to move up through regular promotion competitions; this, together with the exceptional relocation evaluations, has allowed access to better remunerative conditions and given many the possibility to present themselves for access to positions of greater responsibility within the Teaching Career program. In Table 10.7, the current distribution of teachers in the Public-School Teaching Career program is shown by scale level along with the percent index of the *Remuneración Íntegra Mensual* (RIM, Monthly Remuneration)[16] assigned to each scale level.

[16] The RIM of the first teaching scale is set by the government and is used as a reference base on which the amount of the other teaching scales is calculated. The corresponding allowances and bonuses are then added to this remuneration (Minedu, 2018).

Table 10.7 Distribution of teachers in the Public-School Teaching Career (CPM) program according to the teaching scale, 2021

Scale	No. of teachers	% of the CPM	% accumulated	% RIM
Scale 1	73,461	31.3	31.3	100
Scale II	59,795	25.5	56.8	110
Scale III	51,317	21.9	78.6	120
Scale IV	27,510	11.7	90.4	130
Scale V	16,964	7.2	97.6	150
Scale VI	5157	2.2	99.8	175
Scale VII	519	0.2	100.0	190
Scale VIII	0	0.0	100.0	210
Total	234,723	100.0	100.0	

Source Data obtained from the Nexus System of Administration and Control of Teacher Placements, April 2021
Author created

The table above shows that at the higher scales, the percentage of teachers decreases. Likewise, approximately four out of every five teachers in the program are in the first three scale levels, while the percentage of teachers in the three upper scale levels (VI, VII and VIII) does not reach 3%. In fact, there are still no teachers on Scale VIII because the service time requirements to enable competitions for it have not yet been met. It is also notable that the remuneration differences between the first scales are 10% and that it is only after scale V that the remunerative jumps are more substantive and attractive.

Castro and Guadalupe (2021) in their analysis of the evolution of teacher salaries in the country in recent decades conclude that, although it cannot be said that teacher salaries are sufficient or high, "they have recovered compared to those of other professionals, both in absolute terms and corrected by hours worked. In the first case, the advantage of the other professionals has been reduced from a situation in which teacher salaries were doubled to one in which the advantage is approximately 30%. In the second case, the distance has disappeared" (p. 342). However, they point out that the public image of the low remuneration received by teachers does not seem to be sensitive to this salary recovery, which, in their opinion, does not help to position the career as an attractive job option and one in which is in the process of recovery.

10.2.3 Teacher Performance Evaluation (EDD)[17]

Passing the Teacher Performance Evaluation is required for permanence in the career. Failure to pass on three consecutive occasions leads to dismissal. This evaluation is also a requirement to appear in competitions for promotion and access to positions.[18]

Due to its complexity, the implementation of the EDD was projected progressively in a five-year plan by levels of basic education (initial, primary and, finally, secondary) and teaching scales (first the teachers of the higher levels and then those of the lower scale levels[19]). Following this scheme, the Minedu started the EDD with the initial education level (preschool) in the first year (2017) evaluating the teachers of the higher scales and in the second (2018) those of the lower ones. It should be noted that, after two years of implementation, with a prolonged teachers' strike in 2017 and in a context of growing political instability in the country in subsequent years, the Five-Year Plan was subjected to repeated reviews, with the Ministry taking the decision not to apply the evaluation corresponding to 2019. In 2020 and 2021, due to the coronavirus pandemic, the primary and secondary EDDs were temporarily suspended. Thus, to date, Peru has only implemented the initial level (preschool) EDD.

The EDD is organized in three-year cycles made up of one ordinary evaluation, for everyone, and two extraordinary ones, aimed at those who fail. Only those who fail three consecutive times, i.e., the three evaluations of the same cycle, are withdrawn from the career program. Additionally, within each of the three assessments that make up the cycle, any teacher who fails the classroom observation has a new opportunity to be observed.

In its design, the EDD presents classroom observation as a central instrument, which is qualified using rubrics that measure fundamentals of teaching practice. The descriptions of the levels of progression of these performances, selected from the Good Teaching Performance Framework, are narrower than those of the Standards, with the EDD rubrics designed to assess observable aspects in any class session in a period of one to two hours. These rubrics were created by Minedu itself with the support of specialized institutions and experts from the region.

In the EDD, the teacher is observed up to three times: first in a diagnostic (and voluntary) way, which allows them to know the evaluation *setting* and receive feedback. The second observation, which is mandatory, is used to obtain a score on the EDD and is carried out by the director of the school certified by the Ministry, who

[17] It is called EDD for its acronym in Spanish of: Evaluación del Desempeño Docente

[18] This requirement is activated only when the EDD is generalized at one level. For example, in Peru, only at the initial level (preschool) has the EDD been developed for all teachers, so it is only in competitions for promotion and access to positions at this level that the applicant is required to have passed the EDD.

[19] This decision was made to facilitate the creation of evaluation committees for lower-level teachers composed of previously approved higher-level teachers.

previously carried out the diagnostic evaluation and provided feedback based on it.[20] The third observation only proceeds when the teacher has failed the previous one and is carried out by two certified external observers, appointed by Minedu. In addition, the teacher is informed of the observation dates, so he/she has a chance to prepare (Fig. 10.4).

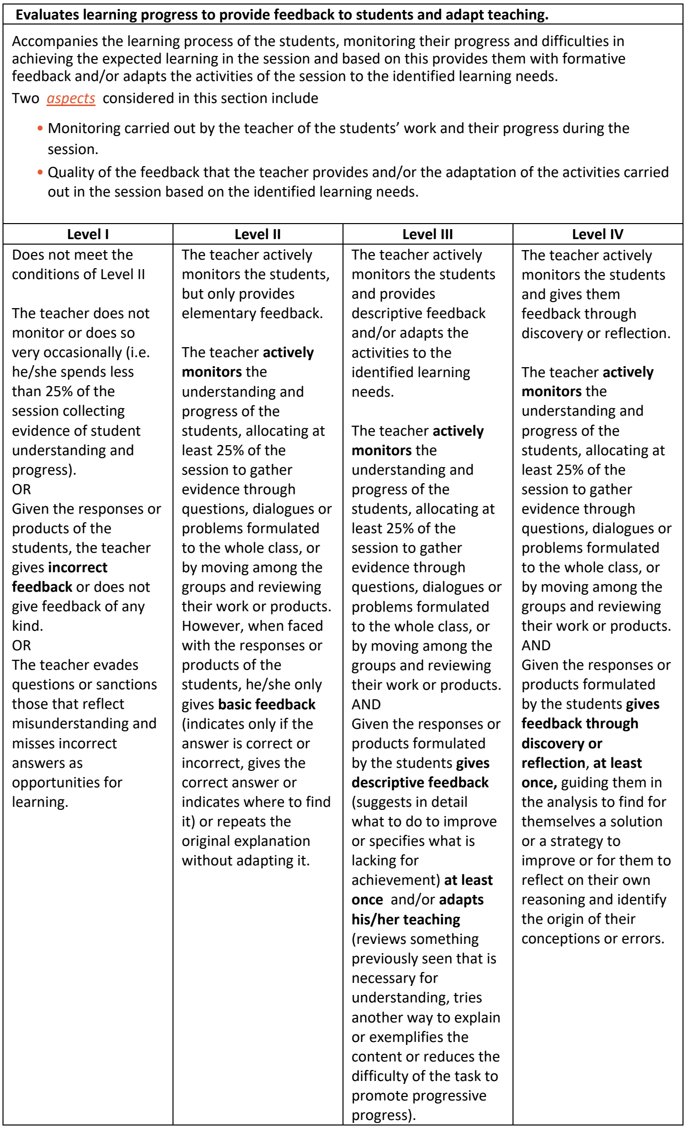
Accompanying the observation of classroom performance is a set of complementary instruments established by Minedu that seek to collect evidence from other sources. These instruments are defined specifically for the level and specialty. Table 10.8 shows the four instruments that cover the 11 performances considered in the EDD for teachers in charge of preschool classrooms with children from 3 to 5 years old. Each performance is assessed with rubrics that describe four levels of achievement, where Level 3 describes the minimum expected performance. Once the performances of the model have been qualified, a simple average is obtained. To pass the evaluation, teachers must achieve an average of 2.6 points, which represents reaching the expected level in most of the performances that are evaluated.

For implementation, Evaluation Committees are formed at the school level with the participation of a director of the institution and two peer teachers from other institutions. These Committees are responsible for applying the instruments, consolidating the results and issuing reports. In multi-grade or single-teacher schools, where there is no school director, the Committee is constituted in the local jurisdiction to which the school belongs with a similar composition.

As can be seen in Table 10.9, the total number of failures of the initial level EDD does not reach 3%. In its original design, the EDD proposed standards *challenging but achievable* for the vast majority of teachers, so as to emphasize its educational purpose: to move toward better practices progressively. A simple evaluation was opted for which would be easy to understand and apply. This was also necessary because its decentralized nature posed difficulties for applying more sophisticated instruments.

Likewise, as a result of successive negotiations with the teachers union, measures were adopted to create the conditions that would allow teachers to show their best performance and minimize the risk of false negatives. These included such things as not evaluating aspects of specific didactic or new curricular approaches, but rather assessing performances related to general pedagogical skills; providing public access to rubrics, instruments and manuals; offering informative workshops to the evaluation subjects and including a diagnostic test to familiarize teachers with the classroom observation instrument, notifying in advance the date of the classroom observation and generating an additional observation by two independent external observers in case of failure during the observation made by the director. In addition, starting at the initial level was selected because this is the level with the least number of teachers and has been the object of more consistent policy interventions, with a focus on curriculum that has been clearer and more sustained.

[20] If these observers are not available, the instrument is given by a specialist or official at the local or regional level who does have the certification.



Evaluates learning progress to provide feedback to students and adapt teaching.

Accompanies the learning process of the students, monitoring their progress and difficulties in achieving the expected learning in the session and based on this provides them with formative feedback and/or adapts the activities of the session to the identified learning needs.

Two *aspects* considered in this section include

- Monitoring carried out by the teacher of the students' work and their progress during the session.
- Quality of the feedback that the teacher provides and/or the adaptation of the activities carried out in the session based on the identified learning needs.

Level I	Level II	Level III	Level IV
Does not meet the conditions of Level II The teacher does not monitor or does so very occasionally (i.e. he/she spends less than 25% of the session collecting evidence of student understanding and progress). OR Given the responses or products of the students, the teacher gives **incorrect feedback** or does not give feedback of any kind. OR The teacher evades questions or sanctions those that reflect misunderstanding and misses incorrect answers as opportunities for learning.	The teacher actively monitors the students, but only provides elementary feedback. The teacher **actively monitors** the understanding and progress of the students, allocating at least 25% of the session to gather evidence through questions, dialogues or problems formulated to the whole class, or by moving among the groups and reviewing their work or products. However, when faced with the responses or products of the students, he/she only gives **basic feedback** (indicates only if the answer is correct or incorrect, gives the correct answer or indicates where to find it) or repeats the original explanation without adapting it.	The teacher actively monitors the students and provides descriptive feedback and/or adapts the activities to the identified learning needs. The teacher **actively monitors** the understanding and progress of the students, allocating at least 25% of the session to gather evidence through questions, dialogues or problems formulated to the whole class, or by moving among the groups and reviewing their work or products. AND Given the responses or products formulated by the students **gives descriptive feedback** (suggests in detail what to do to improve or specifies what is lacking for achievement) **at least once** and/or **adapts his/her teaching** (reviews something previously seen that is necessary for understanding, tries another way to explain or exemplifies the content or reduces the difficulty of the task to promote progressive progress).	The teacher actively monitors the students and gives them feedback through discovery or reflection. The teacher **actively monitors** the understanding and progress of the students, allocating at least 25% of the session to gather evidence through questions, dialogues or problems formulated to the whole class, or by moving among the groups and reviewing their work or products. AND Given the responses or products formulated by the students **gives feedback through discovery or reflection**, **at least once,** guiding them in the analysis to find for themselves a solution or a strategy to improve or for them to reflect on their own reasoning and identify the origin of their conceptions or errors.

Fig. 10.4 Example of classroom observation rubric[21]
Source Evaluación Docente. (s.f)

Nevertheless, these measures did not seem to be sufficient to explain such a high rate of passing the EDD. A more detailed analysis of the results on each instrument and performance evaluated shows that the complementary instruments tended to

[21] In a 60-min session, the teacher must allocate a minimum of 15 min to monitor the students' understanding and progress.

Table 10.8 Instruments and performances evaluated in the EDD—initial level, 2018

Type of instrument	Source	Performance
Observation rubrics	Certified Observer (School Director or external evaluator from the Local Educational Management Units or Minedu)	Actively engages the students to learn
		Promotes reasoning, creativity and critical thinking
		Evaluates progress to adapt teaching and provide feedback
		Promotes an environment of respect and proximity
		Formatively manages the behavior of the students
Checklist	School director or teacher peer	Manages the space to promote learning and well-being
		Manages materials to promote learning and well-being
Survey of satisfaction	Families of students for whom the teacher is responsible	Communicates with families
		Knows and attends to the needs of students
Evidence-based assessment guideline	School officials	Plans teaching and learning processes
		Fulfills role in the school with responsibility and commitment

Author created

Table 10.9 Results of the performance evaluation, 2017–2018

Evaluation	Year	Evaluated	Passed	Passed (%)
Evaluation of teaching performance of initial level (preschool) Section I (higher scales)	2017	5437	5399	99.3
Evaluation of teaching performance of initial level (preschool) Section I (lower scales)	2018	15,831	15,399	97.3
Total		21,268	20,798	97.8

Source EDD 2017–2018 Teacher Evaluation
Author created

raise the teachers' averages. For example, the survey applied to families in 2017, places 99% of the teachers evaluated in performance levels III and IV, i.e., in the competent and outstanding levels, which allows us to presume that the parents were, overall, not particularly demanding. Other complementary instruments evidenced similar behavior.

On the classroom observation instrument, on the other hand, a certain level of variability in the results can be observed, particularly in the more pedagogical rubrics in which greater teaching skills are required, such as promoting higher cognitive skills (reasoning, creativity and critical thinking) or generating high-quality feedback during the learning session. In each case, for performance to be considered at level III, it was sufficient that at least one didactic interaction took place in which these processes were clearly stimulated. Despite the favorable evaluation conditions, we found that in these rubrics there was a significant percentage of teachers who failed to generate a single interaction of this nature during the observed session, as shown in Fig. 10.5.

In conclusion, a set of factors seems to be behind the surprising results of the initial level EDD. Some factors are related to the evaluation model being intentionally simple to achieve standardized observation and application due to the high consequences associated with its outcome. Highly favorable conditions were used so that teachers could show a “ceiling” effect, and the evaluation was directed at the initial, more consolidated level of the system. Other unforeseen events and issues related to possible biases in the application of the instruments could have had an effect and should be investigated to be corrected in future editions.

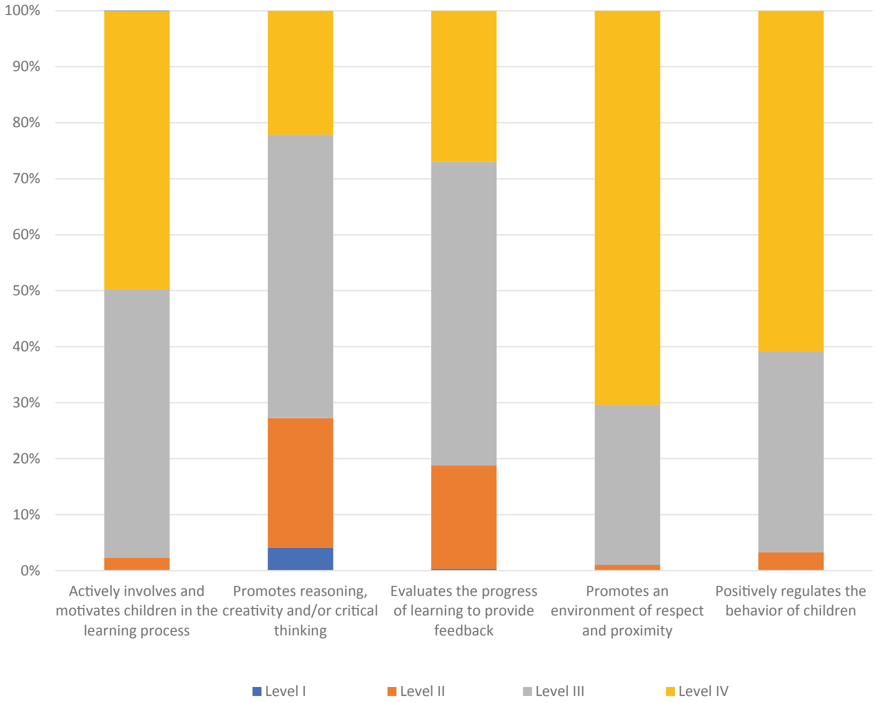


Fig. 10.5 Results of kindergarten teachers (3-5 years) in classroom observation, EDD 2018. *Source* Evaluación Docente. (s.f)

10.3 Overview and Challenges of the Peruvian Teacher Evaluation System

During the 2014–2019 period, Minedu managed to conduct all the evaluations considered in the Teacher Reform Law with criteria of technical quality and transparency. The results have contributed to establishing an orderly teaching career progression based on professional merit, with better remuneration and valued both by the teaching staff and by society as a whole. However, the weaknesses of components of the teaching policy, especially those referring to initial and in-service training, as well as their lack of connectedness, have not allowed the results of the teacher evaluation to be used to nurture the capacities of teachers and achieve the expected level of professionalization. Indeed, there is no systematic evidence on the use of teacher evaluations in the professional development of teachers.

The fact that teacher evaluations have been done efficiently in the period mentioned does not mean the process has been without difficulties. Thus, as a consequence of implementing teacher evaluations, a series of problems have come to light that show the challenges confronting the system not only in terms of improving the evaluations themselves—whose management over time has become more complex—but also above all to put them at the service of professionalizing teaching to guarantee the right to a quality education for the students. Some of these challenges are mentioned below.

10.3.1 Improving Management of the Public-School Teaching Career System

Since the enactment of the Teacher Reform Law, the teacher evaluation system has made it possible to establish a teaching career program, which has unified all civil service teachers in the public sector under a single regime. This in turn has allowed the remuneration, benefits and functions to be systematized in terms of the positions and areas of development contemplated in the career path. In the first two years (2014 and 2015), five exceptional evaluations were implemented that allowed an orderly transition to the new career framework. These evaluations were designed to locate the permanent teachers on the new teaching career scales, to resolve the situation of teachers with provisional credentials, and to fill the existing school leadership positions under the new rules. Once the transition had been made, the implementation of the regular evaluations provided for in the Teacher Reform Law has fulfilled the administrative functions of determining the appointment, promotion and access to positions through meritocratic competitions in which the transparency of the process has been ensured to thus combat, to a large extent, corruption problems that had been associated with these selection processes in the past (Cuenca, 2020; CNE, 2019).

Although the enormous progress in the organization of the Public-School Teaching Career system is undeniable, the implementation of the Teacher Reform Law, and particularly of the teacher evaluation system, has revealed the urgent need to modernize and integrate the information systems on teachers, payroll and placements. These decentralized management systems do not have interoperability or verification mechanisms *online* and in real time that allow access to accurate and reliable information in a timely manner, which frequently leads to irregularities or errors that require attention, adaptations and even specific regulations to be resolved in the midst of the evaluation process.

The absence of an integrated teacher information system forces committees and school officials to dedicate a good part of their time to operational matters such as verifying the requirements and professional trajectory of the teachers participating in the evaluations. To ensure that the system is concentrated on the substantive non-operational aspects of the evaluation, such as the evaluation of pedagogical practices, it is essential to build an integrated information system that allows not only the planning and orderly development of the processes, but also the timely access to information by the teachers themselves. This will not only result in the relief of the operational burden necessary for the implementation of the evaluation system throughout the country and the Minedu itself, but will also provide for greater control, transparency and efficiency of all the processes of the Public-School Teaching Career regime with the consequent trust and legitimacy that this entails.

10.3.2 Building Participatory Processes into the System

In this phase of the teacher evaluation system, the teacher evaluation processes have involved the participation of tens of thousands of teachers and school officials in the processes of implementation, certification and application of instruments. However, the evaluation designs, with some specific exceptions[22] have been the responsibility of the Minedu technical teams, who have defined the instruments, protocols and cut-off points. Thus, the participation of teachers and other actors in the system has been limited in the designs of the models for admission, promotion and performance evaluations in many cases due to the haste with which they had to be created.

As the subject of evaluations is at the first level of the political agenda, their degree of exposure is high and often requires greater social legitimacy and public communication than other technical projects in the sector. To legitimize the system and at the same time make it more relevant to the different contexts, it is essential that the actors in the system be involved in the design of the evaluation proposals for the next phase, particularly the designs of the decentralized stages of the competitions for admission and promotion, as well as performance evaluations by schools and sub-national levels of the education system.

In order to advance this effort further, it is also essential to communicate the results of the evaluations to the actors in the system. Not only the evaluation designs

[22] This has been the case above all in the admission and performance evaluations in positions where consultation with actors such as school directors and provincial and regional authorities have been carried out regarding the designs.

and instruments but also the foundations and assumptions on which they are based should be shared, explicitly connecting them with the framework of Standards to then collect feedback and support that would enrich the models and diversify them. Furthermore, it is essential to take stock of the implementation processes, collecting recommendations from the evaluation committees and of those who participate in the teams that execute these processes.

Unlike what happened at the beginning of the Teacher Reform Law's implementation, at present the positions in the Public-School Teaching Career (CPM) regime are held by teachers who have accessed them through merit. These teachers can be given greater autonomy to manage some processes under their responsibility such as the definition of the instruments and the types of evidence-based information to be considered in the decentralized stages. In addition, it is possible to create consultation spaces for teachers who have reached the higher levels of the CPM and who can contribute with their wisdom to enrich the system.

10.3.3 Institutionalizing and Consolidating the Evaluation System

For the Peruvian teacher evaluation system to be institutionalized and consolidated, it is essential to shore up its autonomy and stability, rethink the quantity and frequency of evaluations to reduce the operational burden and place the emphasis on the dissemination and use of results, as well as on opening a component of research and validity studies that will allow the designs to be improved with evidence-based information.

On the first point, to ensure the autonomy and stability of the evaluation system, it is necessary to recognize that the management of the teacher evaluation system has a high level of technical, logistical, legal and political complexity. Massive and complex evaluation processes with such important consequences for the key actors—the teachers—require sufficient maturity, planning and design times to achieve the expected educational impact. Peru has changed ministerial management eight times since the beginning of the implementation of evaluations set in motion through the Teacher Reform Law. With management changes and institutional problems, emerging demands on the system have been produced that undermine planning and generate stress.

To their credit, the evaluation departments tend to operate within bureaucratic systems at a faster pace than the rest of the offices since their activities are linked; they manage with unpostponable deadlines and usually receive public attention. It is therefore important to review the institutional arrangement that will make the system sustainable and allow for its institutionalization. As the National Education Council ((CNE), 2019, p. 26) points out, "the implementation of the different evaluations established by the Teacher Reform Law constitutes a challenge due to the limitations that the country has in the construction of tests, the presence of institutions and evaluation specialists, and logistic and contractual challenges specific to public management. In this framework, it is pertinent to discuss the possibility of having an independent evaluation institute that seeks to specialize existing human resources for evaluation tasks and to optimize the logistic and budgetary aspects they entail."

As for the second aspect, *rethinking the quantity and frequency of evaluations*, the full implementation of all types of teacher evaluations, including those of teacher performance and for access to other positions, has led to saturation at all levels of the system: school, local, provincial, regional and national. According to the CNE (2019), as time passes, the Teacher Reform Law's evaluation agenda becomes more complex to manage for the Minedu and, especially, for the decentralized educational management bodies that participate in these processes and are overloaded with its activities. Thus, the territorial authorities have been overcome with responsibilities for the implementation of the decentralized stages of the evaluation competitions and performance evaluations. In addition to reducing the operational burden with the integration of information systems, it is necessary to consider reducing the number of evaluations per year, providing sub-national entities with personnel to assume operational management and leaving the tasks of evaluations that require expert judgment such as observing classroom performance for the specialists and pedagogic leaders. Likewise, it would be useful to study the possibility of reducing the number of national competitions and migrating to qualifying tests that have a validity period greater than that of a single competitive exam. Freeing the system of this operational load will contribute to dedicating more to analysis and use of the results.

Finally, consolidating the system will obligate it to be subjected to the research necessary to gauge its technical validity and provide feedback to improve its instruments and processes. Although the Ministry of Education's *Dirección de Evaluación Docente* (DIED, Teacher Evaluation Department) carries out internal processes of validation and performance analysis of its items, a pending task is to direct and publish studies of validity and reliability of the evaluation instruments and designs that are used along with studying the consequential validity and association or relationship of teachers' results with other relevant variables such as student learning outcomes. The latter needs to be planned into the design of the evaluations and organized in collaboration with the efforts of other agencies that collect and process such information. It is essential to stimulate alliances with universities and research centers, as well as to generate incentives for the use of information by local, national and international researchers. Likewise, it seems necessary for Minedu itself to have a research unit that works prospectively and develops a study agenda that allows the designs to be refined.

10.3.4 Placing the Evaluation at the Service of Strengthening the Teachers

The current regulatory framework explicitly establishes that the evaluations that are part of the Public-School Teaching Career (CPM) program have an essentially educational purpose. The information it generates should serve to promote continuous improvement of the teachers and their promotion and mobility through the different areas of job performance within profession.

A positive aspect of the teacher evaluations, implemented within the framework of the career program, has been the feedback from the results. The teachers evaluated have received reports with their individual results highlighting their achievements and areas for improvement. In addition, for those teachers who will be evaluated with the Teacher Performance Evaluation, they can access the rubrics beforehand to learn about the progressions, discuss them with their peers and use them to reflect on their own teaching practices. As for the school administrators and education specialists at local and regional levels, they receive the results of their own evaluations and, in addition, prior to their participation in evaluation committees, they are trained and certified to observe classroom performance. This is a useful tool not only in the context of teacher credentialing exams and performance evaluations, but in general to develop their pedagogical leadership role and to direct mentoring and monitoring processes in their schools. In this way, for the first time in Peru, the majority of school directors have been trained in the use of a pedagogical tool of this nature.

However, although these initiatives have been important, it is necessary that as they are consolidated, the system of teacher evaluations focuses more decisively on strengthening teacher capacities. The teacher evaluations should clearly connect and gain meaning with teacher training and professional career processes, and this connection should be evident. Within this framework, more sophisticated evaluation models could be created to collect information relevant for teacher training and curricular implementation. Thus in this way, evaluation designs could be based on performance standards and be more connected with training institutions for the analysis of results.

One aspect that deserves reflection is the role played by the Teacher Performance Evaluation. As seen in the previous section, due to its characteristics, this evaluation has the greatest potential to provide feedback on teaching practice. However, addressing both its formative and regulatory nature in terms of continuity in the profession generates multiple challenges for its design, implementation and political sustainability. In that area, it would be important to evaluate the possibility of making the EDD independent of the restrictions imposed by the heavy consequences that can result from its application. This would allow its current design to be rethought so that, on the one hand, it could be more linked to the set of professional competency standards, and, on the other hand, it could generate more rigorous evaluation processes with qualified judges for those teachers who show poor performance and those who should not continue in the profession.

In sum, it is essential that teacher evaluations in this new stage are not limited to satisfactorily fulfilling their administrative function, but that they also provide feedback to the system to generate improvements. The evaluation models must be reviewed in the current context, posing questions that lead the way to a fourth phase of the evaluation system, in which the formative function is emphasized and placed at the heart of the designs.

Final reflection

In this chapter, the teacher evaluation policies in Peru have been detailed. The progress and achievements of the evaluations are significant and their impact on building a teaching profession based on merit unquestionable; however, their implementation has revealed a series of challenges that must be addressed by the overall policies regarding teacher development through a complete teacher development policy so that its results are fully taken advantage of to nurture the capacities of teachers and achieve their professionalization.

One of the key factors for the success of evaluation policies is that they are part of comprehensive reforms. The analysis carried out shows that the absence of a comprehensive teaching policy in Peru is leading to an over-demand on the evaluation system, with the risk of confusing the tool with the purpose. Overcoming this issue means strengthening and better coordinating the different components of teaching policy, especially those referring to initial and professional in-service training and development.

Finally, as for the teacher evaluation system itself the design of some of its evaluations needs to be rethought in light of the results and findings that have been obtained after this phase of implementation of the Teacher Reform Law. Specifically, it is necessary to review the credentialing process in the Public-School Teaching Career program, generating alternative evaluation models to fill the positions with suitable professionals in more vulnerable or complex contexts such as those in more remote rural areas. Such designs should start by asking whether the basic skills required to be evaluated are the same as in monolingual urban contexts and proposing exclusive competitions for those positions or additional phases to promote their coverage with good candidates who have not obtained a position but who have passed the entire evaluation. As for the performance evaluations, strengthening their legitimacy and political sustainability by giving priority to their formative function is essential. In this case, it is important to increase the depth and complexity of the design, using the standards to better distinguish among teachers who have high, intermediate and poor performance and to generate distinct routes for strengthening teacher capacities accordingly.

References

Bertoni, E., Elacqua, G., Marotta, L., Martínez, M., Méndez, C., Montalva, V., Olsen, A. S., Santos, H., & Soares, S. (2020). El problema de escasez de docentes en Latinoamérica y las políticas para enfrentarlo. Nota Técnica No IDB - tn - 01883. BID.

Castro, M. P., & Guadalupe, C. (2021). Una mirada a la posición social del docente peruano: remuneraciones, jornada laboral y situación de los hogares. En: C. Guadalupe (Ed.), *La educación peruana más allá del Bicentenario: Nuevos rumbos*. Universidad del Pacífico.

Consejo Nacional de Educación [CNE]. (2017). *Proyecto Educativo Nacional. Balance y Recomendaciones 2016–2017*. CNE.

Consejo Nacional de Educación [CNE]. (2019). *Evaluación del Proyecto Educativo Nacional al 2021*. CNE.

Consejo Nacional de Educación [CNE]. (2020). *El Proyecto Educativo Nacional al 2036: el reto de la ciudadanía plena.* CNE.

Cruz-Aguayo, Y., Hincapié, D., & Rodríguez, C. (2020). *Profesores a prueba: claves para una evaluación docente exitosa.* BID.

Cuenca, R. (2012). *¿Mejores maestros? Balance de políticas docentes 2010–2011. Insumos para el Diálogo 11.* USAID.

Cuenca, R. (2020). *La evaluación docente en el Perú.* IEP.

Cuenca, R., & Vargas Castro, J. C. (2018). *Perú: El estado de políticas públicas docentes.* Diálogo Interamericano - IEP.

Dirección de Formación Docente en Servicio. (2017). *Estándares de Desempeño Docente. Documento de trabajo.*

Evaluación Docente. (s.f). Evaluación en cifras. Ministerio de Educación. Disponible en: https://evaluaciondocente.perueduca.pe/evaluacion-en-cifras/

Evaluación Docente. (s.f). Evaluación del Desempeño Docente Nivel Inicial - Tramo II: Rúbricas de Observación de Aula. Ministerio de Educación. Disponible en: https://evaluaciondocente.perueduca.pe/rubricas-de-observacion-de-aula/pdf/rubrica3-jardin.pdf

Guerrero, L. (2011). *Marco de buen desempeño docente. Documento para la discusión encargado por el Consejo Nacional de Educación.*

Herrero, J. (2012). La carrera magisterial en el contexto actual. *Intercambio, 21*, 1–2.

Lynch, N. (2006). *Los últimos de la clase. Aliados, adversarios y enemigos de la reforma educativa en el Perú.* Centro de Producción Fondo Editorial. Universidad Nacional Mayor de San Marcos.

Ministerio de Educación del Perú [Minedu]. (2013). *Marco de Buen Desempeño Docente.* Ministerio de Educación del Perú.

Ministerio de Educación del Perú [Minedu]. (2018). Ley No 2994. Ley de Reforma Magisterial, Reglamento de la Ley de Reforma Magisterial D.S. No 004-2013-ED y modificatorias. http://www.minedu.gob.pe/reforma-magisterial/pdf-ley-reforma-magisterial/normas-complementarias-de-la-ley-de-reforma-magisterial.pdf

Ministerio de Educación del Perú [Minedu]. (2019). Norma que regula el Concurso Público de Ingreso a la Carrera Pública Magisterial 2019 y que determina los Cuadros de Mérito para la Contratación Docente 2020 - 2021 en Instituciones Públicas. https://evaluaciondocente.perueduca.pe/media/11550870436RVM-N%C2%B0-033-2019-MINEDU.pdf

Ministerio de Educación del Perú [Minedu]. (2020). Estadística de la Calidad Educativa. Censo Escolar 2020. Disponible en: http://escale.minedu.gob.pe/magnitudes

Ministerio de Educación del Perú [Minedu]. (2021). *Resultados de la Encuesta Nacional a Docentes de Instituciones Educativas Públicas de Educación Básica Regular. ENDO Remota 2020. [Diapositivas de Power Point].*

Piscoya, L. (2005). *Cuánto saben nuestros maestros. Una entrada a los diez problemas cardinales de la educación peruana.* UNMSM, Fondo Editorial de COFIDE.

Secretaría de Planificación Estratégica. (2007). *Informe de Resultados de la Evaluación Censal de Docentes de Educación Básica Regular.* Documento de trabajo.

Vázquez, M. del Á., Cordero, G., & Leyva, Y. E. (2014). Análisis comparativo de criterios de desempeño profesional para la enseñanza en cuatro países de América. *Revista Electrónica "Actualidades Investigativas En Educación," 14*(3), 1–20.

Part IV
Teacher Evaluation Systems Around the World: Europe

Chapter 11
Teacher Evaluation in Portugal 12 Years Later: Critical Issues and Possible Directions

Maria Assunção Flores and Eusébio André Machado

Abstract This chapter focuses on the analysis of teacher evaluation in Portugal implemented since 2008 drawing upon a selection of empirical studies carried out between 2012 and 2020. The discussion of the legal framework of current teacher evaluation system in Portugal as well as its evolution over time is also included. In total, 74 studies were reviewed. Findings point to a number of critical aspects related to the absence of an organizational, professional and scientific legitimacy, to the lack of preparation of the evaluators and to the tensions between agency and control inherent in what can be identified as an internal school-centered model. Another important feature emerging from existing research literature is associated with peer evaluation which has been marked by collegiality and collaboration versus competitiveness and individualism. The chapter ends with the discussion of the main findings and their implications for research and policy in the context of teacher evaluation.

11.1 Introduction

Teacher evaluation has been subject to a number of reforms worldwide in an attempt to raise the standards of teaching and to improve the quality of teachers, including their professional development and career advancement. Issues of accountability and growing pressure to increase student achievement within the context of testing regimes and international assessments have also been associated with policy initiatives related to teacher evaluation (Flores, 2010a). Martinez et al. (2016) identified, among other issues, variations in terms of how different teacher evaluation

M. A. Flores (✉)
Research Center on Child Studies, Institute of Education, University of Minho, Braga, Portugal
e-mail: aflores@ie.uminho.pt

E. A. Machado
Portucalense University, Oporto, Portugal
e-mail: eusebio@upt.pt

J. Manzi et al. (eds.), *Teacher Evaluation Around the World*,
Teacher Education, Learning Innovation and Accountability,
https://doi.org/10.1007/978-3-031-13639-9_11

systems in different countries operationalized good teaching as well as the degree of standardization of the classroom observation process.

Existing literature points to different logics highlighting the tensions between formative—oriented toward professional development—and summative purposes—linked to accountability and managerial decisions (Avalos & Assael, 2006; Avidov-Ungar, 2018; Chow et al., 2002; Flores, 2012a, 2018; Stronge, 2006) which are also discussed in the Portuguese teacher evaluation system. Discussing policy and practice of teacher evaluation in Portugal and in the USA, Flores and Derrington (2017) identified three key dimensions to enhance the quality and success of a teacher evaluation policy: the existence of supportive school structures for teacher evaluation; the need to interpreting and managing policy in context while dealing with its mediating factors; and the relevance of the formative dimension of evaluation, particularly the role of supervision.

A number of questions in discussing teacher evaluation have to be considered: Who defines and how the frame of reference for teacher evaluation? Who makes decisions about the procedures and instruments for the evaluation process? Who are the evaluators and what kind of status do they have in the evaluation process? What is the role of teachers in the evaluation process? What kinds of intended and unintended effects may be anticipated? What about ethical issues?

Thus, both the content of the teacher evaluation system and the context in which the system will be used have to be taken into account if it is to be effective and successful (Peterson & Comeaux, 1990). In this paper, we look at teacher evaluation in Portugal drawing upon a review of studies and the examination of the legal framework.

11.2 Setting the Scene: The Portuguese Education System

Despite its geographical condition as a European country, Portugal has gone through difficulties in overcoming its peripheral status presenting recurring indicators of a relative delay when compared to other countries. As far as education is concerned, the processes of schooling and literacy developed at a rather low pace when compared with the majority of European countries.

Although Portugal has been one of the first countries to institute compulsory education (1835), the truth is that, in the 1970s, only 2 in 3 children were sent to school and a quarter of the Portuguese population was illiterate. Since the 1970s, but mainly after the Carnation revolution which occurred in 1974, education has been seen one of the priorities in terms of public policies. Later, after joining the Economic European Community in 1986 (today known as European Union), Portugal has initiated a process of acceleration of schooling of the Portuguese population. At the same time, some of the critical issues of the education system started to get solved as is the case of retention and school dropout.

In this context, the approval of the Fundamental Law of Education (*Lei de Bases do Sistema Educativo*) in 1986 constitutes the turning point of education in Portugal.

Issues of teacher recruitment and education, the qualification of schools and the assurance of better conditions for teaching through policies of support and compensation of disadvantaged children were tackled with more investment in such aspects.

Currently, the education system is organized according to various cycles of study whose duration varies. Elementary education is composed of three cycles: first cycle (corresponding to primary education) comprises four years (pupils aged 6–9); second cycle which includes year 5 and 6 (pupils aged 10–11); and third cycle comprising three years (pupils aged 12–15). In turn, secondary education includes three years—10, 11 and 12 (students aged 16–18). It is important to note that compulsory education entails 12 years, including both elementary (nine years) and secondary education (three years).

Finally, higher education occurs in both polytechnics and universities and comprises cycles of study including *Licenciatura* degree (three years) (Bachelor), Master degree (two years or one and half year) and Ph.D. (three to four years). It is noteworthy that in order to become a teacher in Portugal a Master degree is required for all entrants from pre-school to secondary school. The teacher education organization model currently in place includes a first degree (*Licenciatura*) on a given subject (e.g., Maths, History, etc.) plus a two-year Master degree in Teaching Maths or History. For pre-school and first cycle education, a first degree in Basic Education is required plus a Master degree (one and half year degree). This consecutive model was implemented after the adoption of the Bologna process, and it was in place for the first time in 2007/2008.

Drawing on 2019 official data (Direção-Geral de Estatísticas da Educação e Ciência, Direção de Serviços de Estatísticas da Educação and Divisão de Estatísticas dos Ensinos Básico e Secundário, 2020), the student population in non-higher education system was 1,613,334 corresponding to a schooling rate of 92.2%. In turn, there were in the public sector 146,992 teachers, a figure that has been decreasing as a result of low birth rates. Private and cooperative sector is important in Portugal, comprising 321,409 students (19.9% of the total Portuguese student population). In the private sector, there were 19,834 teachers, most of which work in pre-school, secondary and vocational education.

In Portugal, the teaching workforce in elementary and secondary education is characterized as follows: a high percentage of female teachers (79.9%); the aging issue (almost half of the teachers are over 50 years old); and a big number of teachers are not integrated into the teaching career yet (24,000, around 16% of the total number of teachers). The last Teaching and Learning International Survey (TALIS) report (OECD, 2019) highlights a "dramatic change" in this regard as there was a significant increase of teachers aged 50 or above in Portugal from 28% in TALIS 2013 to 47% in TALIS 2018.

11.3 Teacher Evaluation in Portugal: An Historical Overview

Historically, the development of teacher evaluation in Portugal was marked by three different periods of time corresponding to three different logics (Machado et al., 2012): (i) the first period includes the dictatorship (*Estado Novo*) up until the Red Carnation Revolution in April 1974 during which an external evaluation model, conducted by school inspectors and rectors (principals), was prevalent; (ii) the period between 1974 and 2007 was characterized by the hegemony of a model based on self-evaluation, leaving behind any external dimension (e.g., inspectors, principals and peers); (iii) the period between 2007 up until now has been marked by an internal evaluation model in which peer evaluation is a key feature along with self-evaluation and a mitigated external logic (see Table 11.1).

Up until the revolution occurred on April 25, 1974, under the period of the dictatorship, despite all the changes that have occurred, teacher evaluation was developed generally according to an external model in light of the principle stating that "the evaluators belong to other external organizations or systems, being independent or neutral, and enabling objectivity and distance in order to guarantee an evaluation

Table 11.1 Diachronic evolution of teacher evaluation in Portugal (expanded and adapted from Machado et al., 2012)

Three Phases in the development of teacher evaluation	Model	Definition of the frame of reference	Evaluators	Procedures and methods	Main goals
1st period: dictatorship (*Estado Novo*) up until the Red Carnation Revolution in April 1974	External	External (Ministry of Education)	Inspectors and rectors (principals)	School visits (classroom observation) Annual mark	Political surveillance Career management
2nd period: between 1974 and 2007	Self-evaluation	Internal (Schools)	Teachers being evaluated	Delivery of a self-assessment/annual report to be submitted to the school administration/management body	Career progression
3rd period: from 2007 up until now	Internal	Internal (Schools) and external (Ministry of Education)	Peers (internal and external evaluators)	Delivery of a self-assessment report and, in some cases, classroom observation (external evaluator)	Improving the quality of the educational service Needs' analysis in terms of in-service education Career management

free of subjectivity and partiality and of intra-organizational conflicts" (Machado et al., 2012, p. 74). As such, during this period, within a logic focusing on the scientific/disciplinary, pedagogical and mainly political control (although not primarily concerned about the "outcomes"), teacher evaluation was conducted by the inspector of the school district and by the rectors (principals) of the schools. Within this external model, in accordance with the dictatorship regime, a hierarchical vertical relationship between the evaluator and the teacher being evaluated was prevalent without any kind of dialogic or formative dimension. As such, teacher evaluation would serve the purposes of control of teachers' work prevailing a compliance and normative perspective.

From April 25, 1974, Portugal underwent a revolution period during which the dictatorship was replaced by a democratic regime. Up until 1986, "the issue of teacher performance evaluation was far from the political agenda of the various governments, as teacher evaluation was perceived as being linked to punishment and control that characterized the autocratic past" (Machado et al., 2012, p. 77). However, in 1986, teacher evaluation gained prominence in the political agenda with the publication of the Fundamental Law of Education (*Lei de Bases do Sistema Educativo* - Law no 46/86 14 October), being understood within the perspective of career progression and accountability. Thus, in the post-revolutionary period, a shift in teacher evaluation occurred moving toward a radical self-evaluation system which "has also been reduced to a rather administrative evaluation procedure without any effect in terms of differentiation" (Pacheco & Flores, 1999, p. 189). The same authors argue that, despite its contingently innovative assumptions, teacher evaluation, whose legislative text regulating it was published in 1998, maintains the purpose of certification and ignores the goal of teacher professional growth and school development. Therefore, it is possible to point to a "rather bureaucratic and routinized evaluation process that did not make teachers accountable for their actions, in so far as being accountable for the purpose of career progress was the sole goal" (Machado et al., 2012, p. 79). In other words, teacher evaluation became a mere formality for career progression purposes with no impact on professional development. It involved the writing up of a self-evaluation "critical report" focusing on the activities developed over the course of a given period of time. The report was to be sent to the administration and management body of the school along with the certification of in-service courses and modules. Criteria to evaluate the report included: (i) teaching duties; (ii) pedagogical relationship with the pupils; (iii) fulfillment of syllabi; (iv) management and pedagogical roles; (v) participation in projects and activities within the educational community; (vi) in-service education and its respective credits (1 credit = 25 h of training); (vii) innovative contributions to teaching and learning; and (viii) studies and published work. Teacher evaluation grades were given by the school administration and management body according to two possibilities: satisfactory or unsatisfactory. Only exceptionally an external evaluation would be carried out: (i) having had unsatisfactory in the internal assessment; (ii) the willingness and requirement from the part of the teacher; and (iii) accessing the eighth stage of the teaching career. This would require the constitution of a team of evaluators both internal and external.

11.3.1 The Current Model of Teacher Evaluation in Portugal

In 2007, taking into consideration the issues mentioned earlier, the Ministry of Education decided to initiate a kind of Copernican Revolution by designing a teacher evaluation system which was paradigmatically different from the one in place up until then. A new Teacher Career Statute was issued (Decree-Law number 15/2007, 19th January). Among other features, it has introduced a "more demanding system for teacher performance evaluation with effects on the development of teachers' career" in order to "identify, promote and reward the merit and to value the teaching activity" (see preamble of the Decree-Law) with effects on career advancement.

Teacher career includes from now on ten stages the duration of which is four years, except stage 5 whose duration is two years. Career progression is directly related to teacher performance assessment and depends on getting a grade of "good" as a minimum grade required for all stages of the teacher career, except for stages 4 and 6. In the case of the stages 4 and 6, it is mandatory to obtain a "very good" or "excellent" grade in order to advance to the next level. If it is not the case, teachers are put on a national list and they have to wait for a vacancy. Besides teacher performance assessment, career progression is also dependent on the compulsory frequency of 50 h of in-service education and training for teachers (INSET), except for stage 5 for which only 25 h of training are required. Career progression also depends on classroom observation in the case of teachers on their probationary year and in the case of teachers in stages 2 and 4 of their careers.

The main intention was to abandon teacher evaluation based solely on self-evaluation through a "document of critical reflection" that all teachers had to do. Notwithstanding the legislative subterfuge resulting from a climate of great contestation and controversy that marked the current model of teacher evaluation in Portugal, it implied several legislative changes up until 2012 (Decreto-Regulamentar 2/2008; Decreto-regulamentar 1A/2009, Decreto Regulamentar no 2/2010, Decreto-Regulamentar no 26/2012), the analysis of which is beyond the scope of this chapter.

The process of simplification and adjustment of the model was due to the continuous and strong controversy and contestation from the part of the teachers since 2008 (see Flores, 2009, 2010b, 2012a). One of the most critical aspects related to the existence of a quota system. This represented the abolition of the automatic promotion of teachers in place until 2007 allowing the access from all of the teachers to the top of the teaching career. In order to end this practice, and especially because it represented a situation that financially was hard to maintain from the part of the state, it became necessary to introduce limitations in terms of teacher career progression.

For the first time, teacher evaluation moved from a mere bureaucratic procedure to include effective consequences for teachers. Differentiation of teacher performance was guaranteed by the setting up from the part of government of maximum percentage of the grades of "very good" (between 20 and 25%) and "excellent" (between 5 and 10%) in each school in light of the total number of teachers and the outcome of school evaluation. This quota system has led to competitiveness and to the deterioration of professional relationships and was subject of a strong negative reaction from the part of teachers and teacher unions. Strikes and demonstrations became recurrent leading the government to adopt legislative changes although the quota system exists up until today (see Flores, 2009, 2010b).

In 2012, the current model of teacher evaluation was introduced through the publication of a new legislative text (Decreto-Regulamentar no 26/2012). Again, a process of simplification and decrease of bureaucracy was assumed in order for teachers to focus on teaching and learning. Thus, evaluation cycles became longer, coinciding with the stages of the teaching career (see above): teacher evaluation occurs now every four years instead of each year. There was also a reduction of the dimensions according to which teachers are to be evaluated. Only three of them were adopted: scientific/disciplinary and pedagogical dimension (60%), participation in the school and connection with the community (20%) and in-service education and professional development (20%).

However, the novelty of the 2012 legislative text consisted of the introduction of two evaluation components (see Table 11.2): an internal component under the responsibility of an internal evaluator (the head of the department or school principal) and an external component under the responsibility of a teacher (external evaluator) from another school teaching the same subject and being in the same stage of the teaching career or above as of the teacher being evaluated, and having experience or training in teacher evaluation or pedagogical supervision.

Within the internal component of the system, the school is supposed to define the frame of reference for teacher evaluation, in accordance with its educational project, as well as the instrument for data collection to be used by the internal evaluator. As such, within the teacher evaluation system, schools are granted a relative autonomy

Table 11.2 Synthesis of the teacher evaluation model currently in place in Portugal

	Evaluators	Dimensions	Frame of reference	Methods and procedures	Instruments	Duration
Internal component	Head of department or school principal	Scientific and pedagogical dimension; participation in the school activities and connection with the community; in-service education and professional development	Goals and objectives of the educational project of the school: parameters approved by the pedagogic council of the school	Analysis of the self-evaluation report	Evaluation instrument approved by the pedagogic council of the school	Throughout the period of time corresponding to the stage of the teaching career (usually four years)
External component	A teacher/peer from another school	Scientific and pedagogic dimension	National parameters	Classroom observation	National evaluation instrument and rubrics	180 min

Source Authors

as long as the general guidelines related to teacher evaluation are taken into consideration, mainly as far as the dimensions to be included. In turn, the external component has a national nature and it is up to the Ministry of Education to define the frame of reference to be used by all external evaluators as well as the instrument for data collection.

The selection of external evaluators is the responsibility of School Association's Training Centers which is a local structure usually within a municipality or intermunicipality whose aim is to organize in-service education and training for teachers (INSET). An external evaluator needs to comply with two conditions: being in stage 4 of the teacher career or above and holding a Ph.D., Master degree or a specialization program on teacher performance assessment or pedagogical supervision or having professional experience in terms of pedagogical supervision. Each external evaluator may evaluate ten teachers maximum, and this role needs to be fulfilled within the 35 h of his/her workload per week without any kind of remuneration.

Classroom observation is to be done by the external evaluators and lasts 180 min divided into, at least, two different moments. In order to do classroom observation, external evaluators use a national instrument for data collection in which they have to register positive and negative aspects in relation to: content/subject, use of Portuguese, didactical aspects and relational matters. To facilitate the process of filling in such national instrument for data collection, a frame of reference does exist which clarifies the indicators to be considered in classroom observation and in the grading of the teachers' performance. Both the Ministry of Education and the School Association's Training Centers support external evaluators and organize training for them (25 h training or seminars 3/6 h each).

While the internal evaluator carries out his/her role in accordance with the goals and objectives included in the educational project of the school and the parameters approved by the pedagogical council of the school, the external evaluator is obliged to use the national parameters (Despacho no 13981/2012), as well the reference models for the instruments and rubrics to use for classroom observation purposes.

Classroom observation is an exclusive competency of external evaluators and it is mandatory in the stages 2 and 4 of the teaching career as well as in the case of teachers in the probationary year, for teachers aspiring the grade of "excellent" and for teacher who received unsatisfactory. Classroom observation is conducted during the last two years of the stage of the teaching career in which the teacher being evaluated is located. The external evaluation only evaluates the scientific and pedagogical dimension. This dimension includes four factors: content, knowledge of the Portuguese language, didactics and relational aspects. The grade given by the external evaluation counts 70% and the remaining 30% come from the grade attributed by the internal evaluator. The final grade is obtained according to ten points scale: excellent (9–10); very good (8–8.9); good (6.5–7.9); satisfactory (5–6.4); and unsatisfactory (1–4.9).

Currently, the system of teacher evaluation may be characterized as follows (Decreto-Regulamentar no 26/2012):

(a) The internal dimension of teacher evaluation, granting schools autonomy for the definition of the frame of reference, the construction of evaluation instruments and the organization of the evaluation process;
(b) The self-regulation principle through peer evaluation based on a mitigated and soft hierarchy, despite the inclusion of the figure of the external evaluator (a teacher/peer from another school);
(c) The existence of classroom observation as the main (and almost only) instrument for data collection about the scientific and pedagogical performance, being mandatory in some key stages of the teaching career;
(d) The co-existence of formative and even supervisory logics with summative and grading logics;
(e) The prevalence of a model both internal and external, although it co-exists with self-evaluation processes that were not abolished.

The configuration of the current Portuguese teacher evaluation system is marked by a number of tensions, some of which intrinsic to the definition of the model itself that are visible in other countries as well. In terms of legitimacy, teacher evaluation is organized around a number of options that reinforce the performance control and increase the accountability process through the outcomes in line with current challenges in teacher professionalism (Sachs, 2016).

On the other hand, it noteworthy the introduction of the principles of selectivity and meritocracy aiming at producing an uneven career characterized by a gradual connection between salary and merit. However, it is also important to note that the current teacher evaluation system includes at least two main distinctive features with emancipatory and formative potential in terms of professional development: a school-centered system and peer evaluation. These aspects are in line with structuring options in terms of teacher professionalism in light of existing literature (see Evetts, 2009; Sachs, 2016) which point to teachers' autonomy and agency, collaboration and discretionary judgment as important features of the teaching profession.

11.4 Empirical Studies Reviewed

This section describes the process of selection of the empirical studies included in this review. The criteria and sources are described as well as the methods for data analysis. This chapter seeks to address the following research questions:

1. What do we know and do not know about the **implementation** of teacher evaluation in schools in Portugal after 2008?
2. What does research conducted in Portugal tell about **peer evaluation** and **classroom observation**?
3. What kinds of **issues need to be examined** in further research?

In order to respond to these questions, a selection of empirical studies carried out in Portugal has been reviewed and analyzed. A search was undertaken in the

Table 11.3 Selected studies

	f
Ph.D. theses	16
Master degree dissertations	49
Papers in academic journals	9
Total	74

national database *Repositório Científico de Acesso Aberto em Portugal* (RCAAP) and in academic journals between 2013 and 2020. The search was limited to this period because (i) the meta-analysis of research conducted between 2008 and 2013 was used as a starting point (Machado & Abelha, 2014; Marcos, 2013), and (ii) the framework for the third cycle of teacher evaluation, which includes both internal and external dimensions, was published in 2012 and remains in place up until today (see section above on the current teacher evaluation system). "Teacher evaluation," "teacher performance assessment" and "teacher appraisal" were used as descriptors for the search. Three main criteria were used: (i) only empirical studies were considered; (ii) studies were carried out in mainland Portugal; (iii) studies were published between 2013 and 2020. In total, 74 studies were included in the present analysis (see Table 11.3).

Table 11.4 shows that most studies are qualitative, followed by mixed-method approach and quantitative studies. Most studies are based on teachers' perceptions gathered mainly through interviews and questionnaires. Empirical work based on observation is scarce as are studies using focus group. Other methods included critical incidents, narratives and workshops. The vast majority of the studies included data collection with teachers (71) and 17 studies also included other stakeholders (such as principals and heads of department).

The selected papers were subject to two types of analysis. A descriptive analysis was conducted through the identification of the focus of each study, characteristics of the sample, the methods and their main findings. In second phase, a content analysis was performed (Ryan & Russell Bernard, 2000) using the analytical framework presented in Fig. 11.1.

Three main categories were used, namely: frame of reference (who defines it? How? How is it used?); evaluators (Who are the evaluators and what kind of status do they have?) and purposes and effects (What is teacher evaluation for? How is it used?).

First, all studies were subject to a detailed and descriptive analysis using a set of dimensions: author and year, focus, methods, participants and main findings. Then, a cross-analysis related to each category was conducted in order to look for patterns that make sense beyond every specific case (Huberman & Miles, 1994), without disregarding the particular features of each study. In order to verify the accuracy of the analysis, a "verification" strategy (Creswell, 1998) was used. The review was undertaken by both researchers through regular exchanges and meetings to check the research process and summary of data as well as its interpretation (Lincoln & Guba, 1985).

Table 11.4 Nature of the studies reviewed, methods and participants

Nature of the studies				Methods						Participants	
Mixed-method	Qualitative	Quantitative	Case studies	Questionnaires	Interviews	Observation	Document analysis	Focus group	Others	Teachers	Other stakeholders
23	29	15	7	41	49	8	13	4	11	71	17

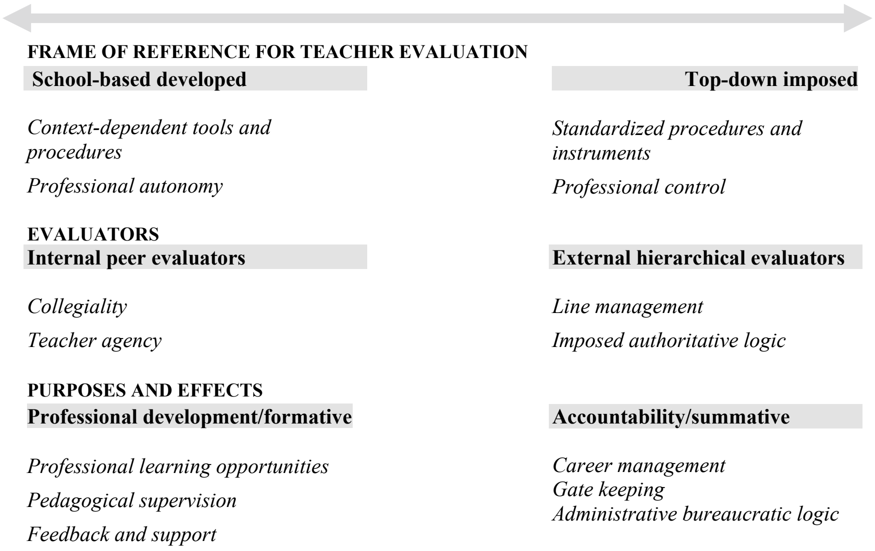

FRAME OF REFERENCE FOR TEACHER EVALUATION
School-based developed
Top-down imposed
Context-dependent tools and procedures
Professional autonomy
Standardized procedures and instruments
Professional control
EVALUATORS
Internal peer evaluators
External hierarchical evaluators
Collegiality
Teacher agency
Line management
Imposed authoritative logic
PURPOSES AND EFFECTS
Professional development/formative
Accountability/summative
Professional learning opportunities
Pedagogical supervision
Feedback and support
Career management
Gate keeping
Administrative bureaucratic logic

Fig. 11.1 Analytical framework. *Source* Authors

Both researchers agreed on the methodological options and procedures with regular verification of all steps in the research process: database search, criteria for selecting the studies, types of analysis to be undertaken, accuracy of the process of analysis and as well as the summary of the findings, their interpretations and organization of final data and conclusions. The main findings are presented in the next section according to the three main research questions and emerging themes.

11.5 Findings

11.5.1 Implementing Teacher Evaluation in Schools: A Disruptive Process

Teacher evaluation in Portugal was marked by the possibility of schools to develop the frame of reference, tools and procedures for the implementation of teacher evaluation. One of the key findings emerging from the review relates to the negative perceptions and experiences of teacher evaluation which are associated with conceptual and processual issues and with the deterioration of school climate and professional relationships (Afonso, 2016; Flores, 2018; Gamero, 2018; Marcos, 2013; Serrano, 2013). In general, studies point to the emergence of negative effects on teachers' work linked to the artificialism of the procedures, the emergence of a climate of competitiveness and conflict and feelings of unfairness and anxiety (Alves, 2013;

Gamero, 2018; Lapo, 2015; Silva & Herdeiro, 2015). In general, teacher evaluation is described as an ineffective process lacking coherence and consistency in regard to the procedures, participants, mainly the evaluators, and outcomes (Soares, 2013). An example of such inconsistencies and lack of coherence related to the mismatch between the formative intentionality of the policy and the bureaucratic nature of its implementation along with a logic of control which was visible in the amount of paper work and extra work for teachers. The effects on teacher learning and growth were rather weak (Soares, 2013); instead teacher evaluation became a rather bureaucratic and administrative process without any practical implications for professional development and improvement of practice.

Lack of motivation and excessive bureaucracy associated with the implementation of teacher evaluation have also been identified (Fialho, 2017; Gamero, 2018; Monteiro, 2015; Silva, 2014a, 2014b). These negative depictions of teacher evaluation stem from the prevalence of a bureaucratic-normative control within a technical-rationality prevalent in the evaluation instruments according to a quantitative logic (Jacinto, 2014) undermining teacher collaboration (Gamero, 2018; Silva, 2014a, 2014b). Added to this is the existence of a great deal of legislative texts which affected school dynamics (Pousada, 2015).

A number of criticisms were associated with the process of implementation of teacher evaluation in schools: the imposed nature of the model, difficulties associated with the evaluators' recruitment, lack of experimentation of the model, resistance from the part of the teachers, lack of specialized training in supervision and the emergence of conflicts and tensions in schools (Coelho, 2015; Duarte, 2015; Monteiro, 2014). The evaluation model is described as unfair and lacking coherence as a result of the quota system which is said to undermine transparency and fairness of the evaluation process leading to feelings of instability and conflict (Antunes, 2014; Serrano, 2015).

Comparing the two legal frameworks of teacher evaluation (issued in 2008 and 2012, respectively), Rola (2014) concluded that no relevant changes were identified in the perceptions of the teachers, both in procedural and content terms, pointing to a rather negative view. Lack of adequate training for evaluators, bureaucracy and the existence of a quota system remain critical features in both periods. Flores (2018) found persisting challenges and perceived effects in a context marked by resistance and rejection of the model of teacher evaluation, namely issues pertaining to the procedures and processes; the role of the participants; the endless legislation and the (unintended) effects on teachers and on schools. An example of such unintended consequences is associated with the deterioration of professional relationships and of school climate which contradicts the rationale of the policy which was based on collaboration and collegiality. The same study showed a number of dilemmas and tensions, among which are: (i) matching the (competing) expectations of both central administration and teachers; (ii) combining short and long-term goals; and (iii) keeping a balance between the summative and bureaucratic requirements and the needs of the teachers.

11.5.2 The Emergence of Tensions Between a Control and a Professional Logic

Clearly, the studies reviewed point to tensions and even contradictions between two different logics. Although issues of teacher collaboration and professional development are identified in the legal framework, a rather technical, bureaucratic and instrumental perspective is prevalent (Afonso, 2016; Cruz, 2013; Jacinto, 2014; Moreira, 2014) leading to a logic of control with no impact on improving practice and fostering teacher professional growth (Alves, 2013). For instance, Dias (2018) found that a summative perspective marked the implementation of teacher evaluation in detriment to the formative approach, emphasizing that the model did not contribute to improve teacher performance nor professional development. The same author concludes that supervisory practices are scarce with no practical effects and are dependent on the profile of the teacher evaluators.

In general, teacher evaluation is used for career management purposes rather than for fostering teacher professional development and improving teaching (Afonso, 2016; Antunes, 2014; Moura, 2014) pointing to the incompatible perspective between formative and summative logics (Coelho, 2015). For instance, Jacinto's study (2014) stresses the prevalence of a normative fidelity logic regarding organizational and supervisory dimensions as well as professional autonomy in face of external and internal regulation. Teachers welcome a more formative dimension but they also recognize that such a dimension is non-existent in practice (Lapo, 2015; Santos, 2017).

Teacher evaluation has been described as a complex and bureaucratic process leading to tensions and conflicts associated with the deterioration of professional relationships among teachers (Marcos, 2013; Serrano, 2015). Overall, the studies show that teacher evaluation did not contribute to improve the conditions for pupil learning and academic achievement nor to the development of an ethos facilitating innovation and professional development (Duarte, 2015; Fernandes, 2014; Fialho, 2017; Marreiros, 2016; Monteiro, 2015; Moreira, 2014; Santos, 2017; Vaz, 2019). The lack of career progression is seen as a factor that hindered the formative dimension of teacher evaluation (Macedo, 2016). Teacher evaluation was not articulated with other evaluation processes in the schools and, thus, it did not contribute to improve schools nor to build relationships and develop communities of professional learning (Pousada, 2015) in most cases linked to the lack of formative and timely feedback (Coelho, 2015).

11.5.3 The Paradox of Peer Evaluation: Collegiality or Lack of Legitimacy?

Peer evaluation is a key feature of the Portuguese model of teacher evaluation. Teacher participation in the evaluation process and its contribution to enhance teacher professionalism and collegial relationships have been advocated since the very beginning of the implementation of the model. Yet, the reviewed studies show that peer evaluation is one of the most critical elements. Negative experiences have been reported pointing to the lack of recognition and legitimacy of peers as evaluators (Duarte, 2015; Monteiro, 2014), as well as feelings of unfairness and anxiety, the existence of a quota system and artificialism (Lapo, 2015). Marcos (2013) identified the prevalence of individual logics with an emphasis on isolation, competitiveness and hierarchy hindering supervision processes and opportunities for teacher professional development.

In addition, the evaluators also claim that they do not enjoy their role as peer evaluators and they report lack of adequate training for the job (Lapo, 2015). Other studies point to the existence of a multiplicity of conceptions and competencies linked to evaluation with an emphasis on the technical-normative ones (related to the compliance with norms and regulations and linear ways of operating) and the prevalence of a regulatory supervision directed toward the attainment of national and local professional standards (which include the monitoring of the teacher evaluation process according to existing internal and external guidelines) (Jacinto, 2014; Silva et al., 2014).

Looking specifically at the role of the internal and external peer evaluator, Queiroga (2016) found that internal peer feedback did not lead to improvement in practice, but it has enhanced collaborative work and reflective practice. He adds that it also contributed to the professional development of the evaluators and the identification of their training needs. Thus, it is possible to identify differences between internal and external evaluators. External peer evaluation was seen as more relevant as external evaluators hold specific training or professional experience in pedagogical supervision thus overcoming the lack of training of the internal evaluators. The same study found that the external evaluator is neutral and the judgment tends not to be influenced by friendship or closed relationship with the teachers being evaluated. Internal evaluators focus on the summative dimension of the evaluation process and do not put into practice their supervisory role (Laranjeira, 2016).

In a similar vein, Vaz (2019) concludes that teachers tend to demonstrate greater acceptance of external and internal evaluators in terms of fairness and impartiality if they come from the same subject and have professional experience and training in pedagogical supervision or evaluation. This view corroborates other studies which point to the need to improve the relationship between evaluators and teachers being evaluated in order to fulfill the requirement of supervising the teaching practice within a positive climate facilitating teacher professional development (Antunes, 2014).

11.5.4 Classroom Observation: Important But Also Critical

Classroom observation was not the main focus of investigation in the reviewed studies. Only eight studies focused on classroom observation as the main topic under investigation (Campos, 2013; Craveiro, 2014; Dias, 2013; Ferreira, 2016; Freitas, 2014; Gomes, 2013; Lopes, 2013; Xavier, 2014). The first six studies found that, while classroom observation is seen as necessary and relevant by both evaluators and teachers being evaluated, its potential is undermined by a number of constraints and problems, namely the lack of adequacy of the instruments used, the ways in which classroom observation is conducted, the lack of the required competencies from the part of the evaluators, the lack of valorization of feedback, the reduced number of lessons observed and the emphasis on the summative dimension (Campos, 2013; Craveiro, 2014; Dias, 2013; Freitas, 2014; Gomes, 2013; Lopes, 2013). The seventh study (Ferreira, 2016) aimed at investigating the variability of data resulting from the implementation of classroom observation guide/rubric (for the scientific and pedagogical component) with physical education teachers. Findings point to the lack of reliability of the instrument for classroom observation along with the variability and dispersion of the data associated with the subjectivity of its use. The eighth study (Xavier, 2014), although recognizing the potential of classroom observation as a strategy for teacher professional development, found that an individualistic and competitive culture emerged which, again, is related to the existence of a quota system. As such, a logic of control and accountability was prevalent which hinders the desired formative dimension.

Other studies do also refer to classroom observation by highlighting, on the one hand, its relevance for teacher professional learning, and on the other hand, the lack of impact on teacher collaboration and professional development. For instance, Alves (2013) found that a supervision perspective through classroom observation did not lead to professional development, corroborating the superficiality and formality identified in other studies (Duarte, 2015). In a similar vein, Dias (2018) conclude that classroom observation is a necessary element in teacher evaluation, but it should be developed within a whole-school project according to the supervision cycle in order to foster collaborative work, self-evaluation and reflection about practice.

As such, teachers tend not to choose classroom observation as part of their evaluation process (see section above for the status of classroom observation in the Portuguese teacher evaluation system) as they disagree with the model (Monteiro, 2014). They refer to issues such as lack of career progression, lack of recognition and legitimacy of the evaluators, lack of adequate training on pedagogical supervision, disagreement in regard to the ways in which supervisors were recruited, and artificialism in classroom observation (Duarte, 2015; Monteiro, 2014).

11.5.5 Raising Awareness of the Relevance of Teacher Evaluation

Despite the overall negative depiction of teacher evaluation emerging from the studies reviewed, some positive features were identified. The potential of teacher evaluation to foster reflection about performance assessment and to stimulate team work was highlighted (Afonso, 2016; Macedo, 2016). Findings point to the awareness of the need for teacher evaluation (Afonso, 2016) and the emergence of a culture of evaluation as a result of learning about evaluation (Jacinto, 2014). For instance, in Lapo's study (2015), although negative aspects were prevalent, some positive features were stressed, namely issues of reflection and improvement when a positive relationship between the evaluator and the teacher being evaluated exists fostering sharing and a good climate. Other studies found that supervision was in general marked by reflection about practice, collaborative work, negotiation of the decision making process and the existence of an interpersonal and democratic relationship between supervisors and teachers (Monteiro, 2014).

11.5.6 Issues that Need Further Research

The review of the studies described earlier clearly shows how the implementation of a new policy on teacher evaluation was, and still is, necessary but also contentious. Issues of artificialism, bureaucracy, individualism and lack of legitimacy of the evaluators were reported. Findings also point to the need for teachers to be more involved in the development of the evaluation process in order to impact practice, professional development and student learning (Campos, 2013; Marcos, 2013). In addition, the need to develop more robust theoretical frameworks for teacher evaluation and classroom observation (Dias, 2018) and to foster supervisory roles of the evaluators (Monteiro, 2015) was also reported. In other words, it is crucial to validate the instruments of data collection, namely for classroom evaluation, but it is also important to clearly identify the dimensions and indicators of teacher performance as well as the conceptions of teaching inherent to the evaluation process. The crucial importance of training for all stakeholders is also stressed in order to develop more sustained and participatory processes of teacher evaluation (Candeias, 2018).

The ambivalent findings shown in some studies in which positive and negative features were highlighted (Afonso, 2016; Lapo, 2015; Macedo, 2016; Monteiro, 2014) are to be related to contextual and personal factors, namely conditions for implementing teacher evaluation in the schools, the supportive role of principals, the belief in the contribution of teacher evaluation to teacher and school development, and the participation of the stakeholders. This is in line with other research. In a study conducted in the USA and Portugal, Flores and Derrington (2017) discussed the problems and strategies developed by principals in both countries to deal with the implementation of teacher evaluation policies. Among other features, the authors identified the need for school principals to balance conflicting goals, to minimize the negative effects of evaluation, to manage tensions of implementation and to make sense of the new policy and its effects at school.

As far as the methodological issues are concerned, the majority of the selected studies report on small-scale investigations, based on convenience samples and on case studies. Additionally, most studies rely on perceptions of the stakeholders, mainly teacher evaluators and teachers being evaluated. It is, therefore, necessary to develop more thorough and systematic research and validation studies in order to get a broader and more consistent picture of teacher evaluation in the Portuguese schools. It would be important to conduct larger studies on how teacher evaluation has been implemented in Portuguese schools and its real effects in terms of job satisfaction, professional development, improvement of practice and its impact on pupil learning and achievement.

As discussed earlier, most studies are based on teachers' perceptions and on small-scale research. It is important to consider other stakeholders including policy makers and school leaders so that the multifaceted and complex dynamic of policy and practice of teacher evaluation (including its micropolitics) may be fully understood. Longitudinal studies are also welcome to gain further insights into the intended and unintended long-term effects of teacher evaluation. To our knowledge, no consequential validity studies have been conducted so far. As such, larger, longitudinal and robust studies, both theoretically and methodologically, are needed in order to evaluate the process but also the impact of the teacher evaluation system in Portugal.

11.6 Discussion and Conclusion

In the OECD report focusing on teacher evaluation in Portugal, a number of issues were identified: the contentious nature of the model; resistance to implementation; difficulties in operationalizing a comprehensive model within a short time span and a number of unintended consequences (see Santiago et al., 2009). The same report highlights, among other features, the need for a balance between improvement and accountability; the need to strengthen teacher evaluation for improvement purposes, providing links between developmental evaluation and career progression evaluation. Twelve years later, these and other issues remain in general unsolved.

The review of studies described in this chapter points to a number of context-specific features of the Portuguese teacher evaluation model, but it also highlights issues that are similar to other contexts. The former is associated with the internal nature of the model—a school-centered model based on peer evaluation—although a mitigated external dimension has been introduced in 2012. The latter relates to the tensions and contradictions of the implementation process, namely the lack of adequacy of the evaluation instruments (see section on current teacher evaluation system in Portugal) and the lack of recognition and legitimacy of the evaluators along with the lack of preparation and training, which have also been reported in other jurisdictions (e.g., Avidov-Ungar, 2018; Lillejord& Børte, 2019; Vaillant, 2008) as well as tensions and even contradictions between formative and summative purposes (Avidov-Ungar, 2018; Clinton & Dawson, 2018). These will be summarized and discussed next.

11.6.1 The Absence of an Organizational, Professional and Scientific Legitimacy

Teacher evaluation in Portugal, in a broader context of the transformation of teacher professionalism, has been implemented according to a centralized and top-down logic without a previous legitimacy process. There was no experimentation phase and the generalization of the system led to tensions, conflicts and unintended consequences (Flores, 2009). Despite the compulsory negotiations with teacher unions, usual strategies to legitimize the process of change were not put into place, namely conducting pilot-studies, the training of the key agents participating in teacher evaluation nor the involvement of experts in developing a theoretical framework. Earlier work has pointed to a process of implementation without experimentation and, as a result, a number of problems related to the evaluation instruments, the profile of the evaluators and the intended and unintended effects (Flores, 2009, 2010b, 2012a, 2018; Machado & Abelha, 2014; Machado et al., 2012). Thus, when the model of teacher evaluation was implemented, it was already marked by a strong absence of organizational, professional and scientific legitimacy, which has led eventually to a strong and almost unanimous contestation involving the Portuguese society. In addition, change has occurred in a sudden way through the shift from a model exclusively based on self-evaluation toward a model based on peer evaluation focused on classroom observation. This represented for the first time in Portuguese schools a massive and systematic process seen by the teachers as excessively intrusive and even violating a space that used to be symbolically marked by feelings of intimacy (the classroom context).

11.6.2 The Lack of Preparation of the Evaluators

The review of the studies points to the impact of the lack of preparation of the stakeholders, namely the evaluators, through a national program of training, although such program occurred a posteriori at a time when contestation had assumed an irreversible dynamic. Huge demonstrations from the part of teachers, including teacher unions and other teacher movements created as a result of the complex and contested implementation of the teacher evaluation model, have been widely documented and highlighted in the media (Flores, 2009, 2010b).

One of the biggest problems in the implementation of teacher evaluation in Portuguese schools related to the lack of professional and scientific legitimacy of the evaluators within a school-centered and peer evaluation model. In face of a flat teacher career, without any hierarchical distinctions, teacher evaluation was developed within an almost totally parity regime without a seniority logic and, thus, conferring a very weak organizational legitimacy to the evaluators.

Along with the weak professional legitimacy, there was no concern about the promotion of the evaluators' scientific and pedagogical legitimacy through a selective process, and in a way a given specialization for the job, but more importantly through the "technical" capacity which is crucial for their function as evaluators and for their recognition from the part of the peers/teachers.

It is noteworthy that, in 2008, when the profound reconfiguration of teacher evaluation in Portugal was initiated, the educational system presented a rather incipient situation regarding supervisory roles, classroom observation and mainly peer evaluation. These features did exist but only in the context of initial teacher education for student teachers, mentors and supervisors in practicum. It was this formative and experiential legacy that was mobilized and adapted to the new reality of teacher evaluation, although difficulties in such process were identified due to the complexity of the process itself that the excessive and unarticulated legislative texts exacerbated. As such, without the recognition from the peers and within the context of the non-existent practices of classroom observation, teacher evaluation has put the educational system and the schools in a state of confusion, division and discomfort with implications for teacher identities, professional relationships and social recognition.

11.6.3 An Internal Model Caught Between Agency and Control

A wide and complex reform such as teacher evaluation occurred in 2008, transforming schools in spaces where all teachers are evaluated and are evaluators, is not compatible with the lack of legitimacy policies. These could have been promoted either via discussion, debate and negotiation processes involving the wider number of professionals as possible (and not only the teacher unions in formal meetings), either through professional development strategies, the training of the stakeholders and the development of pilot-experiences to validate both technically and politically the intended reform of teacher evaluation. None of this was taken into consideration in the Portuguese case.

Albeit teacher professionalism has been marked in recent years by greater performativity and control (Flores, 2012b; Sachs, 2016), the demands for participation, engagement and agency remain to be seen as hallmarks of the teaching profession. In the Portuguese context, the strong technical-rationality logic of the implementation process, and more importantly the voluntarist belief from the part of the Ministry of Education that no legitimacy strategies were needed, gave rise to intense reactive dynamics with effects on the entire Portuguese society and to a kind of blocking of teacher evaluation. As Sachs (2016) stresses, teacher professionalism is shaped by external environments, and, in times of increased accountability and regulation, it is possible to identify various discourses, gaining legitimacy and impact on how professionalism is understood and enacted.

Paradoxically, although the policy has been designed and developed within a centralized and top-down logic, the teacher evaluation model adopted in Portugal in 2008 and notwithstanding the changes introduced over the years, has assumed a clear component of school autonomy and teacher agency. The Portuguese model conferred a wide responsibility to each specific context/school to conceive, manage and implement teacher evaluation, namely as far as the evaluators, the frame of reference and the instruments for data collection were concerned. The vast majority of the problems identified in the reviewed studies focus exactly on the impact of this political option of reinforcing the internal nature of evaluation which has led, collaterally, to profound effects on the *ethos* and personal relationships in the schools (Flores, 2012a, 2018).

The political narrative of the justification of teacher evaluation based on the principle of improving teacher performance was rapidly questioned as a result of the emergence of competitiveness, individualism and lack of trust. Thus, teacher evaluation has led to a strong impact on professional and school cultures undermining the collegial motivation around educational projects and the teaching and learning process. As Kyriakides and Demetriou (2007, p. 45) argue, "power and conflict can be considered core operational concepts that capture the essence of the field of politics with regard to teacher evaluation."

Thus, the aspect that would be at first sight a potentially formative feature of the model has turned out to be one the main reason for contestation and conflict. Teacher autonomy and agency is a principle of paramount importance, and even the hallmark of the teaching profession, both theoretically and practically, but it is not compatible with ingenuous voluntarisms from the part of the stakeholders, mainly policy makers and school principals, without an adequate regulation, especially a pedagogical, scientific and professional regulation. In other words, although autonomy and agency are features of the existing model, clear theoretical and methodological frameworks are needed if the implementation of teacher evaluation system is to be successful.

In fact, the political option of transferring to schools a high responsibility in the implementation of teacher evaluation gave rise to a state of confusion originating multiple frames of reference and a confusion around the procedures to be developed. These have led to question the teacher evaluation system itself and the importance of teacher evaluation in terms of professional development. Schools have become places characterized by a highly conflicting climate and by a loss of professional and organizational *ethos*, leading to bureaucratic cultures that, in turn, became obstacles to the implementation of teacher evaluation. As Braun et al. (2011, p. 585) suggest, "policies are intimately shaped and influenced by school-specific factors, even though in much central policy making and research, these sorts of constraints, pressures and enablers of policy enactments tend to be neglected."

11.6.4 Peer Evaluation: Collegiality and Collaboration Versus Competitiveness and Individualism

Peer evaluation is one of the most interesting aspects of the Portuguese case, not only due to its relative originality, but mainly due to the lessons learned in this regard as such a strategy is more consistent with an emancipatory, collegial and empowered perspective of the teaching profession. Yet, the Portuguese case shows that a regime marked by a total parity, based on a career without any kind of differentiation, entails serious problems that need to be discussed.

The reviewed studies point to the lack of recognition of the peers/teachers in relation to the role of the evaluators and the associated effects on the school climate, reinforcing, paradoxically, strategies of competitiveness and individualism. In this regard, the Portuguese case illustrates that formal legitimacy, via legislative texts, is not enough to confer legitimacy to an evaluator in face of his/her peers. The main criticism documented in the reviewed studies deals with the lack of scientific and pedagogical legitimacy, both from the part of evaluators and those being evaluated. The educational system (basic and secondary education), since 1974, has developed a strong parity culture, but the implementation of a teacher evaluation system with effects on career progression demands mechanisms of organizational, scientific and pedagogical legitimacy, even considering the "professionalization" of the evaluator/supervisor in each school, as it is the case of other educational systems.

Peer evaluation, along with the school responsibility for the definition of the frame of reference as well as the production of instruments for data collection, has caused a number of problems which are well documented in the reviewed studies, particularly bureaucracy and the lack of equity in the system as it was possible to identify positive and negative experiences in schools as a result of their own resources, dynamics and organizational capacity. This implies the need for a clear national frame of reference that may function within a regime of subsidiarity with the local frames of reference. The lack of a national frame of reference (although it existed during the second cycle of teacher evaluation—2009 to 2011) has left teachers without a clear orientation has led to an increase of bureaucracy as a defense strategy and has brought about a generalized perception of lack of objectivity and fairness in light of the different procedures, documents and support instruments that schools had generated themselves.

In addition, peer evaluation has led to a strong feeling that it was at the service of career management, professional control and pressure for performativity. This situation has jeopardized the formative and pedagogical dimension of teacher evaluation. As such, the potential of peer evaluation was highly undermined as classroom observation did not meet the desire for improvement, regulation and professional collaboration. Once again, the reviewed studies show the incompatibility, in the same teacher evaluation system, of summative and formative purposes, of career management and professional development, and in other words, of the desire for control and emancipatory narratives.

References

Afonso, R. (2016). *A avaliação de desempenho docente vista por professores: realidades, expectativas, desafios e oportunidades.* Unpublished doctoral dissertation, University of Évora, Portugal.

Alves, M. (2013). *Avaliação do desempenho docente e supervisão pedagógica: um estudo de caso num agrupamento de escolas.* Unpublished doctoral dissertation, University of Aveiro, Portugal.

Antunes, M. (2014). *Desenvolvimento profissional dos professores: perspetivas sobre a avaliação do desempenho docente.* Unpublished Master Degree Dissertation, University of Évora, Portugal.

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.

Avidov-Ungar, O. (2018). Teacher evaluation following reform: The Israeli perspective. *Quality Assurance in Education, 36*(4), 511–527.

Braun, A., Ball, S. J., Maguire, M., & Hoskins, K. (2011). Taking context seriously: Towards explaining policy enactments in the secondary school. *Discourse: Studies in the Cultural Politics of Education, 32,* 585–596.

Campos, S. (2013). *O lugar da observação de aulas na avaliação do desempenho docente: que contributos para o desenvolvimento profissional dos professores?* Unpublished Master Degree Dissertation, University of Minho, Portugal.

Candeias, P. (2018). *Perceção dos principais agentes envolvidos no processo de avaliação de desempenho docente.* Unpublished Master Degree Dissertation, University of Beira Interior, Portugal.

Chow, A. P. Y., Wong, E. K. P., Yeung, A. S., & Mo, K. W. (2002). Teachers' perceptions of appraiser/appraise relationships. *Journal of Personnel Evaluation in Education, 16*, 85–101.

Clinton, J., & Dawson, G. (2018). Enfranchising the profession through evaluation: A story from Australia. *Teachers and Teaching, 24*(3), 312–327.

Coelho, M. (2015). *Avaliação de desempenho docente: efeitos no desenvolvimento profissional.* Unpublished Master Degree Dissertation, Lisbon School of Education/IPL, Portugal.

Craveiro, C. (2014). *Impacto da observação de aulas na avaliação de professores muito experientes.* Unpublished Master Degree Dissertation, University of Lisbon, Portugal.

Creswell, J. W. (1998). *Qualitative inquiry and research design: Choosing among five traditions.* Sage.

Cruz, M. (2013). *Avaliação do desempenho docente: estudo exploratório sobre a perspetiva dos avaliadores.* Unpublished Master Degree Dissertation, Higher Institute of Education and Sciences, Lisbon, Portugal.

Decree-Law number 13981/2012, 26th October.

Decree-Law number 15/2007, 19th January.

Decree-Law number 1-A/2009, 5th January.

Decree-Law number 2/2008, 10th January.

Decree-Law number 2/2010, 23th June.

Decree-Law number 26/2012, 21th February.

Dias, P. (2018). *Supervisão Pedagógica e Desenvolvimento Profissional na Avaliação de Desempenho Docente: Perceções de Avaliadores e Avaliados.* Unpublished doctoral dissertation, University of Coimbra, Portugal.

Dias, P. A. (2013). Observação de aulas em contexto de ADD: função classificatória ou emancipatória da classe docente? *Gestão e Desenvolvimento, 21*, 289–304.

Direção-Geral de Estatísticas da Educação e Ciência, Direção de Serviços de Estatísticas da Educação & Divisão de Estatísticas dos Ensinos Básico e Secundário. (2020). *Educação em Números - Portugal 2020.* Direção-Geral de Estatísticas da Educação e Ciência.

Duarte, A. (2015). *Avaliação de desempenho docente e seus atores.* Unpublished Master Degree Dissertation, Lisbon School of Education/IPL, Portugal.

Evetts, J. (2009). The management of professionalism. A contemporary paradox. In S. Gewirtz et al. (Eds.), *Changing teacher professionalism. International trends, challenges and ways forward* (pp. 19–30). Routledge.

Fernandes, M. H. (2014). *Desenvolvimento profissional docente e avaliação de desempenho: perceções de professores experientes.* Unpublished Master Degree Dissertation, University of Coimbra, Portugal.

Ferreira, R. (2016). *Avaliação do desempenho docente em Educação Física: análise da variância dos resultados obtidos através do sistema de observação de aulas.* Unpublished doctoral dissertation, University of Trás-os-Montes e Alto Douro, Portugal.

Fialho, A. P. (2017). *A avaliação do desempenho docente na escola atual: da legislação à operacionalização.* Unpublished Master Degree Dissertation, Catholic Portuguese *University.*

Flores, M. A. (2009). Da avaliação de professores: reflexões sobre o caso português (On teacher evaluation: Reflections from the Portuguese case). *Revista Iberoamericana De Evaluation Educativa, 2*, 239–246.

Flores, M. A. (2010a). *A avaliação de Professores numa Perspectiva Internacional: Sentidos e Implicações* (*Teacher evaluation from an international perspective: Meanings and implications*). Areal Editores.

Flores, M. A. (2010b). Teacher performance appraisal in Portugal: The (im)possibilities of a contested model. *Mediterranean Journal of Educational Studies, 15*, 41–60.

Flores, M. A. (2012a). The implementation of a new policy on teacher appraisal in Portugal: How do teachers experience it at school? *Educational Assessment Evaluation and Accountability, 24*, 351–368.

Flores, M. A. (2012b). Teachers' work and lives: A European perspective. In C. Day (Ed.), *The Routledge international handbook of teacher and school development* (pp. 94–107). Routledge.

Flores, M. A. (2018). Teacher evaluation in Portugal: Persisting challenges and perceived effects. *Teachers and Teaching, 24*(3), 223–245.

Flores, M. A., & Derrington, M. L. (2017). School principals' views of teacher evaluation policy: Lessons learned from two empirical studies. *International Journal of Leadership in Education, 20*(4), 416–431.

Freitas, M. (2014). *Avaliação do desempenho docente: observação de aulas no 1.º ciclo do Ensino Básico na região autónoma da Madeira.* Unpublished Master Degree Dissertation, Open University, Portugal.

Gamero, M. (2018). *Implicações da avaliação do desempenho docente na aprendizagem dos professores e no seu desenvolvimento profissional: perspetivas de professores de inglês.* Unpublished doctoral dissertation, University of Évora, Portugal.

Gomes, C. (2013). *Perceção dos professores face à observação de aulas, em contexto de avaliação do desempenho.* Unpublished Master Degree Dissertation, Higher Institute of Education and Sciences, Lisbon, Portugal.

Huberman, A. B., & Miles, M. (1994). Data management and analysis methods. In N. K. Denzin & Y. S. Lincoln (Eds.), *Handbook of qualitative research* (pp. 228–244). Sage.

Jacinto, M. (2014). *Esferas de influência na avaliação de professores: das políticas avaliativas às conceções e práticas de avaliação numa escola básica e secundária.* Unpublished doctoral dissertation, University of Lisbon, Portugal.

Kyriakides, L., & Demetriou, D. (2007). Introducing a teacher evaluation system based on teacher effectiveness research: An investigation of stakeholders' perceptions. *Journal of Personnel Evaluation in Education, 20*, 43–64. https://doi.org/10.1007/s11092-007-9046-3

Lapo, M. (2015). *Formação e avaliação de desempenho: contributos para o desenvolvimento profissional*. Unpublished doctoral dissertation, University of Minho, Portugal.

Laranjeira, M. (2016). *O papel da supervisão na componente interna da avaliação docente e o seu contributo para o desenvolvimento profissional: estudo de caso numa escola secundária.* Unpublished Master Degree Dissertation, University of Lisbon, Portugal.

Law number 46/86, 14th October.

Lillejord, S., & Børte, K. (2019). Trapped between accountability and professional learning? School leaders and teacher evaluation. *Professional Development in Education, 46*(2), 274–291.

Lincoln, Y. S., & Guba, E. G. (1985). *Naturalistic inquiry*. Sage.

Lopes, C. (2013). *Supervisão e avaliação de desempenho docente: perceções dos professores sobre observação de aulas*. Unpublished Master Degree Dissertation, Lusófona University, Porto, Portugal.

Macedo, M. (2016). *Avaliação do desempenho docente enquanto orientadora do desenvolvimento profissional*. Unpublished Master Degree Dissertation, *Polytechnic Institute of* Castelo Branco, Portugal.

Machado, A., Abelha, M., Barreira, C., & Salgueiro, A. (2012). Avaliação pelos pares: Percurso normativo da avaliação do desempenho docente em Portugal. *Revista Portuguesa de Pedagogia*, *46*(I), 73–93.

Machado, P. D. E. J., & Abelha, P. D. M. (2014). Avaliação de Professores: que lições do caso português? *Olhares: Revista Do Departamento De Educação Da Unifesp*, *2*(1), 55–80.

Marcos, A. (2013). *Lógicas de supervisão pedagógica em contexto de avaliação de desempenho docente*. Unpublished doctoral dissertation, Portucalense University, Portugal.

Marreiros, C. (2016). *A influência da avaliação no desempenho docente do educador de infância*. Unpublished Master Degree Dissertation, Lisbon School of Education/IPL, Portugal.

Martinez, F., Taut, S., & Schaaf, K. (2016). Classroom observation for evaluating and improving teaching: An international perspective. *Studies in Educational Evaluation, 49*, 15–29.

Monteiro, G. (2015). *Supervisão promotora de mudança e inovação na avaliação do desempenho docente: estudo de caso*. Unpublished Master Degree Dissertation, *Polytechnic Institute of* Castelo Branco, Portugal.

Monteiro, S. (2014). *Avaliação do desempenho docente e desenvolvimento profissional dos professores de línguas-culturas: o papel do supervisor pedagógico*. Unpublished doctoral dissertation, University of Minho, Portugal.

Moreira, M. (2014). *As organizações educativas e as políticas de avaliação do desempenho profissional docente: as consequências pessoais, profissionais e organizacionais da avaliação do desempenho profissional docente*. Unpublished doctoral dissertation, University of Minho, Portugal.

Moura, A. (2014). *O contributo da avaliação do desempenho docente para o desenvolvimento profissional: a perspetiva dos diretores e dos docentes avaliadores e avaliados*. Unpublished Master Degree Dissertation, Catholic Portuguese *University*.

OECD. (2019). *TALIS 2018 results (Vol. I): Teachers and school leaders as lifelong learners*. TALIS, OECD Publishing. https://doi.org/10.1787/1d0bc92a-en

Pacheco, J. A., & Flores, M. A. (1999). *Formação e avaliação de professores*. Porto Editora.

Peterson, P., & Comeaux, M. A. (1990). Evaluating the systems. Teachers' perspectives on teacher evaluation. *Educational Evaluation and Policy Analysis, 12*, 3–24. https://doi.org/10.2307/1163584

Pousada, M. (2015). *Avaliação de desempenho docente: Contributos para as práticas de avaliação na escola*. Unpublished doctoral dissertation, University of Minho, Portugal.

Queiroga, L. (2016). *Avaliação do desempenho docente: contributo da avaliação pelos pares para o desenvolvimento profissional dos professores*. Unpublished doctoral dissertation, University of Coimbra, Portugal.

Rola, M. (2014). *Dois modelos de avaliação de desempenho docente: perceções dos seus protagonistas*. Unpublished Master Degree Dissertation, University of Évora, Portugal.

Ryan, G. W., & Bernard, H. R. (2000). Data management and analysis methods. In N. K. Denzin & Y. S. Lincoln (Eds.), *Handbook of qualitative research* (2nd ed., pp. 769–802). Sage.

Sachs, J. (2016). Teacher professionalism: Why are we still talking about it? *Teachers and Teaching Theory and Practice, 22*(4), 413–425. https://doi.org/10.1080/13540602.2015.1082732

Santiago, P., Roseveare, D., Van Amelsvoort, G., Manzi, J., & Matthews, P. (2009). *Teacher evaluation in Portugal* (OECD report).

Santos, A. (2017). *A avaliação de desempenho docente: Registos de um percurso seguido um estudo de caso no 1º CEB.* Unpublished Master Degree Dissertation, University of Porto, Portugal.

Serrano, É. (2013). *Contributos da supervisão pedagógica no âmbito da avaliação de desempenho docente.* Unpublished Master Degree Dissertation, Open University, Portugal.

Serrano, N. (2015). *Perceção dos professores face à avaliação e face ao modelo de avaliação do desempenho docente.* Unpublished Master Degree Dissertation, *Polytechnic Institute of* Castelo Branco, Portugal.

Silva, A., & Herdeiro, R. (2015). Avaliação do desempenho docente: conflitos, incertezas e busca de sentido(s). *Educar Em Revista, 1*, 137–156.

Silva, A., Machado, M., & Leite, T. (2014). Avaliação de desempenho docente, supervisão e desenvolvimento profissional. *Da Investigação Às Práticas, 5*(1), 41–66.

Silva, I. R. (2014a). *Formação para avaliação do desempenho docente. Perceções dos avaliadores do grupo disciplinar de Artes Visuais – 600.* Unpublished Master Degree Dissertation, Higher Institute of Education and Sciences, Lisbon, Portugal.

Silva, J. P. (2014b). *A avaliação de desempenho docente: implicações no processo de ensino-aprendizagem.* Unpublished Master Degree Dissertation, University of Minho, Portugal.

Soares, R. (2013). *Avaliação de desempenho docente: perceção dos professores e implicações na sua prática.* Unpublished Master Degree Dissertation, Open University, Portugal.

Stronge, J. H. (2006). *Evaluating teaching: A guide to current thinking and best practice* (2nd ed.). Corwin Press.

Vaillant, D. (2008). Algunos marcos referenciales en el evaluación del desempeño docente (Framework for teacher evaluation). *Revista Iberoamaricana De Evaluacion Educativa, 1*, 7–22.

Vaz, A. (2019). *Avaliação do desempenho docente: sentidos e desafios na perspetiva de professores.* Unpublished doctoral dissertation, University of Évora, Portugal.

Xavier, M. (2014). *Observação de aulas no contexto da avaliação de desempenho docente e desenvolvimento profissional.* Unpublished Master Degree Dissertation, Lusófona University, Portugal.

Chapter 12
Promoting Teaching Quality Through Classroom Observation and Feedback: Design of a Program in the German State of Baden-Württemberg

Evelin Ruth-Herbein, Julia Larissa Maier, and Benjamin Fauth

Abstract Teaching quality is positively associated with student outcomes such as achievement and motivation. Thus, in teacher education and training, a variety of assessment tools are used to provide (pre-service) teachers with feedback on their teaching in class. However, these instruments vary regarding their psychometric quality, and there is no common theoretical and empirical basis for the formative assessment of teaching quality across different types of schools. Consequently, the project "Promoting Teaching Quality through Classroom Observation and Feedback" was initiated in the German federal state of Baden-Württemberg. It includes, as a core element, the development of an observation form for external observations, targeting cognitive activation, student support, and classroom management via eleven items. The assessments based on the form are used to provide feedback in the context of teacher training/teacher education and peer feedback. This chapter, first, embeds the present project in the political context of Baden-Württemberg. Second, it describes the theoretical background underpinning the conceptualization of the observation form as well as the accompanying materials and workshop. Third, it presents the scientific studies planned in connection with the project: from the pilot study to widespread use in practice. Initial results are reported and discussed.

E. Ruth-Herbein (✉) · J. L. Maier · B. Fauth
Institute for Educational Analysis (IBBW), Stuttgart, Germany
e-mail: evelin.ruth-herbein@ibbw.kv.bwl.de

J. L. Maier
e-mail: julia.maier@ibbw.kv.bwl.de

B. Fauth
e-mail: benjamin.fauth@ibbw.kv.bwl.de

B. Fauth
University of Tübingen, Tübingen, Germany

J. Manzi et al. (eds.), *Teacher Evaluation Around the World*,
Teacher Education, Learning Innovation and Accountability,
https://doi.org/10.1007/978-3-031-13639-9_12

12.1 Introduction

In this chapter, we describe an approach to classroom observations that has been conceptualized to promote teaching quality through formative feedback. Traditional high-stakes teacher evaluations have been criticized as ineffective for teachers' professional development. Consequently, in the German federal state of Baden-Württemberg, an approach was chosen that uses teacher evaluations to promote professionalization and professional development by feeding back state-of-the-art classroom observations to teachers.

In the following, we will first give an insight into the political context in which the evaluation system will be implemented (Sect. 12.2). One reason why the project was launched is that the large German federal state of Baden-Württemberg has experienced a significant drop in student test scores in recent years. Because teaching quality has been identified as a potential lever to increase student performance, the project entitled "Promoting Teaching Quality through Classroom Observation and Feedback" was initiated as a cooperation project of the Institute for Educational Analysis Baden-Württemberg (IBBW) and the Center for School Quality and Teacher Education (ZSL) in the German state of Baden-Württemberg in 2019. The former institution is responsible for the conception of the observation tool and scientific monitoring, the latter for the accompanying support system. Both institutions are subordinated to the state's Ministry of Education.

Second, we describe the characteristics of the evaluation system, including the theoretical model on which it is based (Sect. 12.3). This model assumes that three basic dimensions of teaching quality are crucial for the development of students' achievement and motivation: cognitive activation, student support, and classroom management. We provide an overview of the items included to assess these three basic dimensions. Additionally, we discuss the relationship between generic and subject-specific aspects of teaching quality and how our observation tool deals with this relationship. We are aware that such a large observation system will not work without a comprehensive support system. In an additional section, we thus describe two professional development workshops that will be put in place to aid the implementation of the project in everyday school practice. One workshop aims to facilitate reliable and valid classroom observations. In a second workshop, participants will learn how to provide effective feedback for teachers based on classroom observations.

In a final section (Sect. 12.4), we provide an overview of several studies that are being carried out to verify the utility and feasibility of the form and manual; we also describe validation studies on the psychometric properties of the instrument. Additionally, we outline the validation agenda for the near future and the evaluation of the instruments during broad dissemination.

12.2 Antecedents of the Teacher Evaluation System

Classroom observations are a direct measure of teaching quality commonly used in many countries to evaluate teacher performance. Observation methods differ substantially regarding the theoretical framework they apply, the instruments they use, and the consequences that follow from the evaluation. Unlike countries such as the United States, Germany does not have a tradition of high-stakes teacher evaluation. Teacher evaluation systems are implemented by governmental institutions at the state level (German *Bundesländer*). The results of classroom observations are usually communicated to schools in the form of whole-school (rather than teacher specific) scores. Oftentimes, evaluations are followed up by a counseling process as part of school development measures (see OECD, 2013; Taut & Rakoczy, 2016). However, this approach to teacher evaluation has recently come up for debate, and policy-makers have started to look for alternative and potentially more effective evaluation systems.

The starting point for the present project and the search for an alternative approach to promoting teaching quality is an observed change in student performance within the past few years. National and international large-scale assessments gained attention in Germany after the publication of the findings of PISA 2000. These studies revealed that German students did not score as highly as most people in Germany had assumed. However, student test scores differed between German federal states. Nationwide comparative studies invariably showed that students in Baden-Württemberg were among the highest performers in the country. This situation changed during the 2010s. A study published in 2017 showed a significant increase from 2011 to 2016 in the proportion of students who did not meet the standards in German (reading: +3.1%, listening: +6.2%) and math (+6.0%). At the same time, the proportion of those with excellent test results decreased (reading: −1.7%, listening: −2.0%, math: −6.0%; Stanat et al., 2017).

In response to this declining student performance in nationwide large-scale standard tests, a new educational quality policy was implemented in the German state of Baden-Württemberg (Ministerium für Kultus, Jugend und Sport Baden-Württemberg, 2017). The state government of Baden-Württemberg developed a quality strategy for primary and secondary education. This policy entailed adopting a more research-based approach to teachers' professional development and to school development programs. Additionally, based on recent research findings, policy-makers identified teaching quality as one key factor influencing student development (e.g., Hattie, 2010). For instance, various empirical studies have found that teaching quality is positively related to student achievement, motivation, and interest (e.g., Fauth et al., 2019; Lipowsky et al., 2009). Thus, the promotion of teaching quality in everyday classroom instruction was regarded as a potential lever to increase student performance in the long run. Within this context, a project entitled "Promoting Teaching Quality through Classroom Observation and Feedback" was initiated. Unlike most other teacher evaluation systems in Germany, this project seeks to provide formative feedback to individual teachers. It hence does not employ teacher evaluation as a high-stakes test but as a measure to promote individual professional development and improve teaching quality in schools.

The project is aimed at teachers in the German state of Baden-Württemberg at all stages of their careers, regardless of the subject, grade level, or school track. In the school year 2019/2020, these were over 110,000 teachers (Statistisches Landesamt Baden-Württemberg, 2020b) in 3548 public schools and 462 private schools (Statistisches Landesamt Baden-Württemberg, 2020a). The feedback given to them by colleagues or advisors should help them to reflect on and develop their teaching quality. The approach developed in this project is to be used within initial teacher education programs and in subsequent programs for professional development. In general, professional development (PD) programs may be designed based on research results obtained from studies examining the effectiveness of teacher training programs (see Darling-Hammond et al., 2017; Kennedy, 2016) and research on how students and teachers learn (Kennedy, 2016). Features of effective PD programs include, for example, feedback and coaching (Darling-Hammond et al., 2017; Lipowsky & Rzejak, 2021).

The aim of the present project is to establish an effective program that focuses directly on teacher action in the classroom and can be broadly implemented across all schools on a repeated and ongoing basis. Therefore, we chose a step-by-step approach that involves repeated studies under increasingly realistic conditions before the program is widely disseminated in practice. To achieve the goal of improving teaching quality, we drew on the formative assessment approach when conceptualizing the program. Formative assessment has been found to effectively promote the development of achievement, motivation, and self-regulation. It incorporates three elements: (i) the determination of a person's performance status, (ii) the interpretation of the results and the provision of feedback derived from it, and (iii) the identification of actions to promote further development (Andersson & Palm, 2018; Black & Wiliam, 2009). These three elements were chosen as underlying core components of the program. To operationalize these elements, we first needed to develop a reliable, valid, and feasible measurement tool for assessing teaching quality. This was the starting point for facilitating effective and constructive feedback.

A variety of different approaches are used to assess teaching quality in the state of Baden-Württemberg. Observation forms that are filled out by external classroom observers are especially used in teacher education. However, these observation forms vary greatly. Furthermore, these tools often lack sound theoretical foundations, and their psychometric properties are almost never evaluated. In addition, their use varies across school types, and even across schools, which makes it difficult to achieve a common, coherent understanding of teaching quality. Based on this situation, and to support quality development in schools across the state, the Baden-Württemberg state government aimed to develop one single evaluation form to be used when assessing teaching quality. This form is to be used within teacher education and teacher training programs, although the goal is not to use this form to make judgments and evaluations.

Because formative assessment is more effective when feedback emerges from the results of the assessment and support steps are derived accordingly (see Andersson & Palm, 2018; Black & Wiliam, 2009) and because feedback and coaching have been found to be effective features of PD (Darling-Hammond et al., 2017; Kraft et al., 2018; Lipowsky & Rzejak, 2021), the "Promoting Teaching Quality through Classroom Observation and Feedback" project also considers these elements. To provide teachers with effective and constructive feedback on their teaching quality, participants in the feedback process should answer three key questions: "Where am I going?," "How am I going?," and "Where to next?" (Hattie & Timperley, 2007; p. 87). To facilitate these subsequent steps, the project developed a sophisticated support system to train external observers in providing personal peer feedback based on their observations, in addition to developing an observation and feedback form.

12.3 Characteristics of the Evaluation System

12.3.1 Theoretical Foundation

During the conceptualization phase, we aimed to develop an instrument that was research-based and enabled formative external ratings of and feedback on teaching quality. Research studies have defined and operationalized teaching quality slightly differently via various indicators and dimensions (Praetorius et al., 2018; Wisniewski & Zierer, 2020). However, in German-speaking countries, a conceptual framework based on three basic dimensions of teaching quality—cognitive activation, student support, and classroom management—is empirically and theoretically well established (Praetorius et al., 2018). This framework suggests that these three dimensions are crucial for students' cognitive and motivational development. These basic dimensions of teaching quality form a potential answer to three questions that are crucial for classroom instruction:

1. To what degree are students stimulated to think about the content and engage in higher-order thinking processes? (cognitive activation)
2. How well does the teacher support students' learning processes? Is everyone in the classroom treated with respect and appreciation? (student support)
3. Are lessons managed effectively such that disruptions are avoided, all students engage with the learning content, and time on task is maximized? (classroom management)

The key features of *cognitive activation* are a clear focus on the relevant content, challenging tasks, and the exploration of concepts, ideas, and prior knowledge (Lipowsky et al., 2009). These classroom practices should foster students' cognitive engagement, which should in turn lead to deeper knowledge and lasting learning (Fauth et al., 2019; Klieme et al., 2009). *Student support* refers to both the cognitive and the socio-emotional aspects of teaching. The cognitive part of this dimension includes individual, positive, and constructive teacher feedback, a positive approach to student errors and misconceptions, and scaffolding for when students struggle with the content being taught (Brophy, 2000; Klieme et al., 2009). The second aspect focuses on caring teacher behavior as well as warmth and respect in the classroom (Fauth et al., 2019). *Classroom management* is a well-known concept in educational research (e.g., Kounin, 1970) that focuses on classroom rules and procedures, strategies for coping with disruptions, and smooth transitions. Effective classroom management provides time on task, which can be seen as a necessary precondition for active engagement in learning (Emmer & Stough, 2001; Fauth et al., 2019). These dimensions are very similar to the three domains conceptualized in the Classroom Assessment Scoring System (instructional support, emotional support, and classroom organization; Pianta & Hamre, 2009). However, there are some differences, particularly in the area of cognitive activation, which has a different focus than the notion of instructional support proposed by Pianta and Hamre (2009). Regarding the effectiveness of the three basic dimensions, cognitive activation and classroom management have been shown to predict cognitive student outcomes (Kyriakides et al., 2013; Lipowsky et al., 2009; Seidel & Shavelson, 2007), whereas a supportive climate was found to be especially connected to students' motivational and interest development (Fauth et al., 2014; Kunter et al., 2013).

All three dimensions have been assessed using a variety of different items and indicators in previous studies (Praetorius et al., 2018; Taut & Rakoczy, 2016). To create an observation form that was practical and easily usable in evaluations, we chose a narrowly defined selection of items. The selection process was guided by two criteria: First, we only included items proven to be positively related to students' outcomes, e.g., achievement, motivation, and/or interest. Furthermore, we chose items that were representative of the corresponding dimension of teaching quality. Second, we designed the observation form based on reliable and valid instruments that have successfully been used in educational research (e.g., Lotz et al., 2013; Rakoczy & Pauli, 2006) as well as on assessment tools that had shown high practical relevance at school. These instruments were used to select the items and to give them a final wording. Finally, the observation form consisted of 11 items: four with regard to cognitive activation, four pertaining to student support, and three referring to classroom management (see Fig. 12.1).

For cognitive activation, we chose one item to assess whether lessons clearly focused on the content and the specific concepts that students need to understand for lasting learning. This focus can be seen as a precondition for the other aspects of cognitive activation: the exploration of students' thinking and understanding through questioning and formative assessments and—based on these explorations—the development of cognitively challenging tasks that teachers use to engage students in knowledge construction and higher-order thinking (Praetorius et al., 2018). The fourth item included in the cognitive activation dimension is student engagement and explicitly focuses on students' behavior; all of the other items focus on teacher behavior.

In the field of student support, one item focuses on the feedback that teachers provide to their students. The indicators for this item explain that effective feedback should be helpful for future learning and should thus focus on the process of solving a task and students' misconceptions in a certain area. The second item—scaffolding—focuses on situations in which students are struggling with understanding and asks

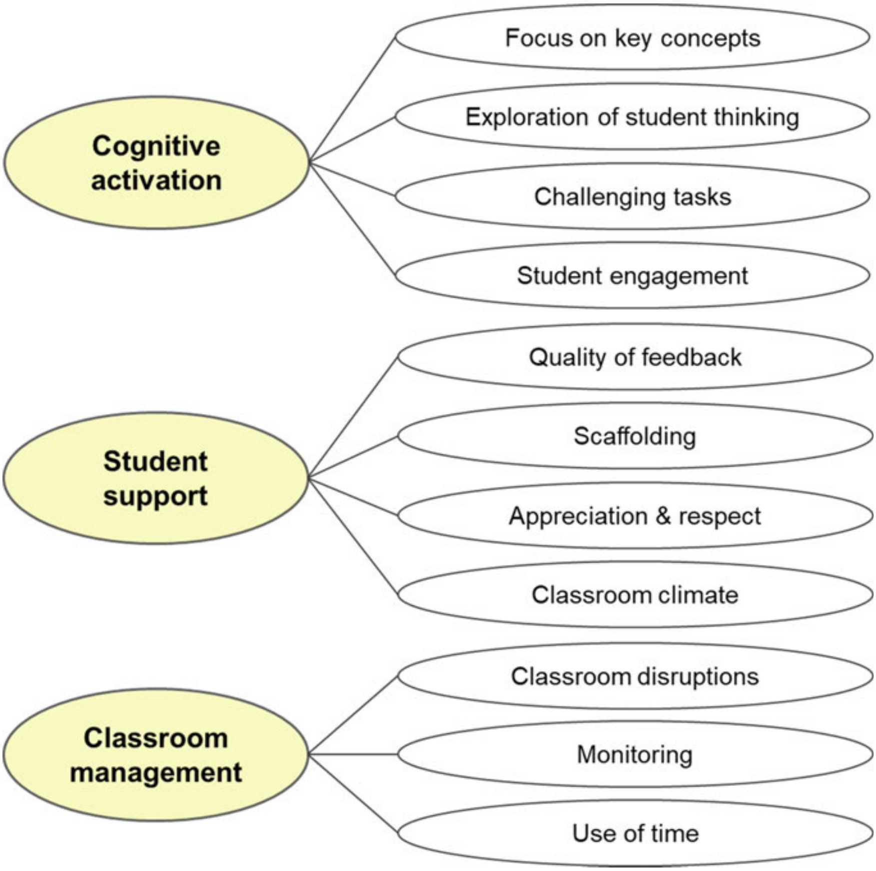


Fig. 12.1 Selected aspects of teaching quality covered by the observation form and their categorization according to the three basic dimensions

about the teacher's ability to provide good support in these situations. Two additional items cover the socio-emotional aspects of student support. One asks about teachers' respect and appreciation for the students' perspective. Another item—the item focusing on student behavior in this dimension—asks whether students treat their classmates and the teacher with respect and appreciation. We call this the "classroom climate."

In the area of classroom management, student behavior plays an important role, too. This is reflected in the "classroom disruptions" item, which focuses on students' classroom discipline and asks about whether students stick to the rules and keep noise in the classroom to an appropriate level. Two further items focus on teacher behavior: "Monitoring" relates to whether the teacher is aware of what a student is doing and whether they are present in the classroom as well as to Kounin's (1970) with-it-ness. "Use of time" asks whether the time available is actually used to engage with the learning content and not to deal with organizational issues or as unnecessary waiting time—during transitions from one phase of a lesson to another, for instance.

12.3.2 The Observation Tool: Generic and Subject-Specific Aspects

At the core of the evaluation system, there is an observation tool consisting of an observation form, a manual including background information and indicators for each item, and a set of domain-specific explanations including specific (video) examples for each domain (see Fig. 12.2). The framework of the three basic dimensions of teaching quality was developed in video observation studies on mathematics instruction and applied to other subject domains (Klieme et al., 2001). Conceptually, the basic dimensions of teaching quality are assumed to be generic in nature. However, there is an ongoing debate in the international literature about how domain-specific aspects are related to more general aspects of teaching quality (Praetorius et al., 2018).

The observation tool we developed in our project adopts an innovative approach to address these issues. As described above, the actual observation form and the manual consist of generic items and indicators. These indicators are then explained in an additional document (the "third level" of the observation tool) with domain-specific explanations and examples. Thus, the observation tool can be divided into three different levels:

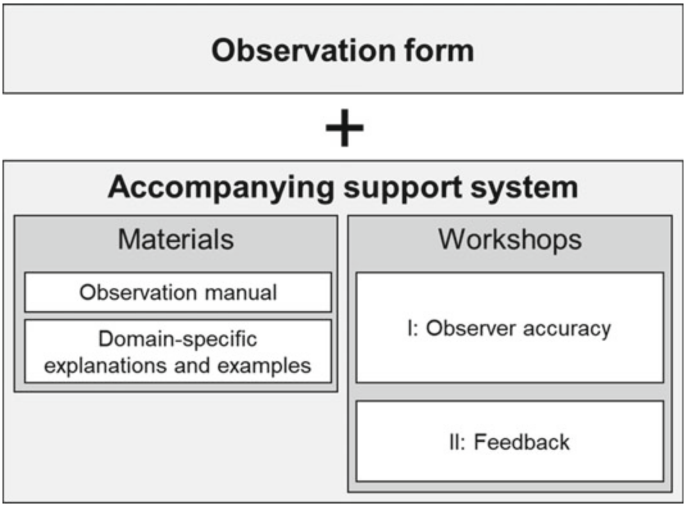


Fig. 12.2 Components of the evaluation system, consisting of the observational form and the accompanying support system

1. The observation form, with the 11 items, a 4-point Likert scale ranging from 1 = *not true* to 4 = *totally true* for each item, and room for observational notes. For an example, see the “challenging tasks” item below.
2. The observation manual, with an introduction to the theoretical foundations underpinning the three basic dimensions (see Sect. 12.3.1) and an introduction to how to use the instrument during classroom observations. Additionally, the manual contains a short description of the theoretical background to each item as well as a set of observable indicators that individuals rating this item will have to consider.
3. While the first two levels are the same for all subjects, grade levels, and school tracks, the third level should include domain-specific rating explanations and examples of what the indicators would look like in a certain subject. Domain-specific examples of classroom interactions are at the core of this third level—they demonstrate what a certain rating would look like in a certain subject. These examples may take the form of verbal descriptions, transcripts, or videos. The domain-specific explanations and examples developed to date have concerned the field of math education. Differentiated materials for all subjects are planned.

In summary, the instrument we use is generic at the first two levels and domain specific at an underlying third level. The domain-specific aspects are not equally important for each item. For example, items like “classroom disruptions” or “appreciation and respect” may take a very similar form in different subjects. Consequently, the third level for those items will look very similar. On the other hand, items like “focus on key concepts” or “challenging tasks” may look quite different in different subjects.

In the following, we will give an example of the “challenging tasks” item. The (translated) wording of the item is: “The teacher uses tasks and questions that challenge students’ higher-order thinking.” The manual provides details about the theoretical foundations for this particular item. This includes the idea that students’ own reasoning about complex relationships between different concepts is an important precondition for lasting learning. This reasoning can be supported by tasks and questions that do not merely require students to reproduce facts. Instead, students should develop their own ideas, analyze relationships between different concepts, and transfer knowledge to new contexts. Additionally, the manual provides several indicators that raters should use to form their judgment. Translated examples of indicators of this item are:

- “The teacher presents tasks and questions that require more than just yes or no answers from students”;
- “The teacher presents tasks and questions that require more than just reproducing previously learned facts or applying clear procedures”;
- “Students are encouraged to develop their own ideas to solve a task.”

These are the first three indicators for this item; there are another six indicators formulated in a similar way. The description of the items in the manual is completed by references to the various research papers and codebooks used to formulate the items and indicators.

Domain-specific explanations and examples developed to date have concerned the field of math education. These explanations include general information on how to use the observation form, the manual, and the item-specific guidelines about typical classroom situations in which an item becomes particularly relevant (e.g., for the "challenging tasks" item, this might be situations in which a teacher introduces new assignments to the students). We collaborated with ten experts in math education (practitioners) to formulate math-specific didactic principles for some of the items. In the case of challenging tasks, these principles included student activities such as problem solving, modeling, and mathematical reasoning. Additionally, the explanations provide video examples of five to ten minutes duration that serve as anchors for each of the four categories of the Likert scale used by raters.

Thus, the domain-specific explanations and examples build a bridge between generic aspects of teaching quality (e.g., cognitive activation: challenging tasks and questions) and very subject-specific didactic principles that have to be explicated for each subject. In the near future, several working groups will start developing domain-specific explanations and examples for further subjects. All of these groups consist of several expert practitioners (expert teachers) as well as researchers working on subject-specific didactics.

12.3.3 The Accompanying Support System

We regard the manual and the above-described domain-specific explanations and examples as part of the support system that was developed to ensure that the observation instrument transitioned smoothly into school practice (see Fig. 12.2 and Sect. 12.4). Regarding the psychometric quality of classroom observations, research shows that intensive rater training is required to correctly use observation tools (Bell et al., 2014; Taut & Rakoczy, 2016). Thus, a professional development workshop was established to ensure that the professionals who used the instrument provided reliable and valid ratings. In the long term, this workshop will be part of initial teacher education programs and of subsequent programs for professional development in Baden-Württemberg. It will thereby be adapted to serve the needs of the specific target groups, for example, teachers, teacher trainers, mentors, or school administrators. This is necessary, because they differ in their expertise regarding the assessment of teaching quality. The concept of the workshop was based on several core components that are assumed to lead to effective teacher training (see e.g., Darling-Hammond et al., 2017; Kraft et al., 2018): theoretical input, opportunity to practice, feedback, and spaces for structured reflection that allowed for extensive discussion between participants. These core components were operationalized via methods such as online seminars (synchronous and asynchronous), ratings of five-minute sequences of video-taped math lessons, feedback on participants' own ratings and comparisons with expert ratings, self-regulatory learning strategy prompts, and small group reflections. Theoretical input was provided in two online seminars. The first seminar covered the theoretical foundations of the concept of teaching quality (see Sect. 12.3.1) and an introduction to the items used in the observation instrument. In an additional online seminar, participants received training in diagnostic skills, diagnostic basics for classroom observations, and typical observation errors. Because of the COVID-19 pandemic, the pilot version of this rater training was conducted completely online. Future studies will show if the same professional development workshop would also work with participants who attend in person or with a blended learning concept.

In an additional professional development workshop, participants will learn how to provide effective feedback for teachers based on classroom observations. This workshop is currently under development. It will be a crucial part of the whole project, because teachers may not benefit from even the best classroom observation if it is not translated into effective feedback for the particular teacher.

12.4 Results and Consequences of the Evaluation System

To develop an observation form that is effective and can be implemented successfully, this project has adopted a stepwise approach (Gottfredson et al., 2015; Humphrey et al., 2016): (1) conceptualization, (2) (practical) feedback on the conceptualization, (3) pilot study, (4) validation, (5) effectiveness study, and (6) broad dissemination and evaluation in practice (see Fig. 12.3). It thus used a symbiotic implementation strategy combining research and practice (Maaß & Artigue, 2013; Parchmann et al., 2006).

Overall, our aim was to prepare the ground for the successful implementation of the observation form in practice and to ensure that it is used in qualitatively excellent ways as the basis for effective feedback. Related to the overall process of the project, the scientific monitoring described in the following focuses on the project's first part, i.e., the development of a sound instrument that allows reliable and valid assessments of teaching quality.

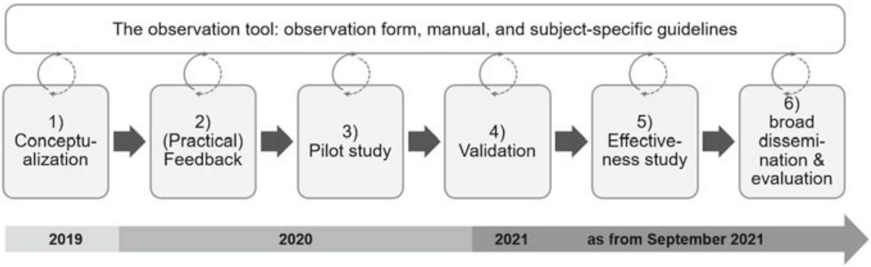


Fig. 12.3 Stepwise approach to develop and evaluate the observation form and the accompanying support systems

12.4.1 (Practical) Feedback on the Conceptualization of the Form and Manual

After conceptualization (Step 1 depicted in Fig. 12.3—see Sect. 12.3 for details), we subjected the derived observation form and its accompanying manual to an initial review (Step 2, see Fig. 12.3). To apply the symbiotic implementation strategy, we included empirical educational researchers as well as practitioners and educational administrators in this process (Maaß & Artigue, 2013). By conducting discussions and soliciting written comments, we received feedback on the theoretical foundations underpinning the form as well as on its fit with the requirements of the school context. Regarding the former, a scientific consortium consisting of experts on teaching quality research assessed whether the selected items covered the relevant aspects of the three basic dimensions of teaching quality. Regarding the latter, the focus was on whether the form met the needs of teachers and persons working in the context of teacher training/teacher education. The questions we used to guide the feedback procedure covered multiple aspects of the implementation process (see Briesch et al., 2013; Gottfredson et al., 2015; Humphrey et al., 2016) and were formulated as open questions. They asked about (i) the comprehensibility of the different elements of the manual, (ii) the scope and depth of the background information provided, (iii) the positive and negative indicators listed, and (iv) the feasibility of the form and manual. Based on the results, the contents and structure of the form and manual were adapted. This included the specification of the items, the selection of the indicators, the scope of the theoretical background, and the graduation of the response scale. With regard to the indicators, the researchers and practitioners commented on the described behaviors as well as on their frequency of occurrence. For example, with regard to the item on scaffolding, the majority of indicators focused on macro-scaffolding. This included the teacher's provision of differentiated or supplementary tasks. To strengthen behaviors related to micro-scaffolding, further indicators were added. One example is: "When queries arise, the teacher explains clearly and understandably." A parallel approach was chosen for the other items and indicators.

12.4.2 Pilot Study

After the revision of the instrument, we conducted a pilot study (Step 3). The aim of the study was twofold. First, we wished to examine the implementation of the observation form and the components of the support system, i.e., the manual and the professional development workshop. The study assessed the perceived utility and feasibility of the form and manual as well as the relative advantages of the form compared to existing observation sheets. Regarding the implementation of the workshop, we assessed utility, relevance, and feasibility. The second goal was to evaluate the psychometric quality of the observation form. We examined whether rater agreement was satisfactory after the workshop.

Ten experts in teacher training/teacher education in the field of math participated in the study. Due to the complexity of the project, we decided to concentrate on one specific subject, i.e., math, during this first empirical step. Nevertheless, future steps will also take other subjects into account. Half of the participants were female. Concerning their teaching qualification, half of the experts held a qualification for lower secondary school, the other half for Gymnasium, the highest school track in Germany. All participants reported being involved in teacher training, meaning that they provided further or in-service training for already-qualified teachers ($M = 11.45$ years; $SD = 6.49$). In addition, seven persons reported working in teacher education, meaning that they provided pre-service training for students who are studying to become teachers ($M = 10.57$ years; $SD = 3.06$). In the pilot study, all experts participated in the newly developed workshops, which took 30 h, including the time used for online surveys and video rating. To ensure high treatment fidelity and standardization, the workshop was offered by the developers of the observation form and workshop. The workshop took the form of e-learning sessions using video conferences, online surveys, and other online discussion formats. We intended the conceptualization of the observation form and workshop to be a cooperative and symbiotic process and thus embedded repeated discussions, focusing on issues like the comprehensibility and feasibility of the form and manual, in the workshop. More precisely, we used the second part of the workshop to revise certain indicators from the manual wherever necessary and to develop domain-specific explanations and examples for math lessons.

To answer the research questions, we assessed implementation of the form, manual, and workshop via questionnaires at repeated measurement points during and after the workshop. In addition, we asked the participants to rate 10 five-minute video clips showing math classes after the workshop. All 10 participants rated the same 10 video clips using the observation form, so that we could assess rater agreement among all raters. To assess rater agreement, we used the average absolute deviation index (AD_M; Burke et al., 1999). For all 11 items of the observation form, the average deviations were below the cut-off value of 0.67 scale points (see Burke et al., 1999). Thus, rater agreement was satisfactory after the workshop. Regarding the implementation of the form and manual, we found high perceived utility in the pilot study (translated item example: “Using the observation form gives me many advantages in assessing teaching quality in practice.”). Concerning feasibility, the participants reported that the elements were highly transferable into practice (translated item example: “I will be directly able to use the observation form to observe teaching quality in class.”). In addition, they reported a relative advantage of the new form over existing instruments (translated item example: “The use of the observation form enables me to focus more precisely on essential and empirically proven aspects of teaching quality than before.”). Concerning the workshop, participants reported high utility, relevance, and feasibility across measurement points in the pilot study (translated item example: “Attending the workshop is worthwhile because you need the content to use the observation form effectively.”).

12.4.3 Validation Study

The pilot study was followed by a validation study (see Step 4 depicted in Fig. 12.3). The goal of the study was again to examine the implementation of the form and manual and to evaluate the psychometric quality of the observation form. Concerning the later, we examined whether rater agreement was satisfactory and tested convergent and predictive validity in a subsequent step. The ten observers trained during the pilot study participated in the validation process. In a kickoff-meeting, they received an introduction to the video material, which was necessary because the videos rated in the pilot study differed from the videos used in the validation study in terms of their length and up-to-date-ness. While 5-min sequences were used for practical reasons within the pilot study, the validation study went one step further into practice. Therefore, whole 45-min sequences were rated, because this is the typical duration of a lesson in German schools. The videos the participants applied their ratings to were existing classroom videos from the Pythagoras study (Klieme et al., 2014). The data from this study includes videos from $N = 34$ classrooms, teaching quality ratings for these classes, and student achievement data. This data will enable us to evaluate interrater agreement as well as the convergent and predictive validity of our newly developed observation form in the future. With regard to the former, the rater agreement for these 45-min sequences was as satisfactory as for the short sequences rated during the pilot phase. To obtain evidence regarding convergent validity, we will compare the ratings of our participants with the existing teaching quality ratings from the Pythagoras study within a multitrait-multimethod analysis. High correlations between items assessing similar aspects of teaching quality would be an indication of high convergent validity. In a second step, the comparison of our participants' ratings with the student achievement data from the Pythagoras study will provide information about the predictive validity of the developed observation form. The data collection and analysis phase for this part of the study is still running. When participants were asked about the utility and feasibility again after validation, the values slightly dropped. One reason might be that, during the validation process, the participants experienced additional challenges when using the form because they were required to use it intensively within a short period of time.

12.4.4 Next Steps: Effectiveness Study and Broad Dissemination

Based on the results of the pilot study and validation, we plan to conduct an effectiveness study next (Step 5). To move one step further into practice, the workshop will no longer be offered by the developers of the observation form, the supplementary material, or the workshop (see pilot study) but by trained pilot study participants who are practitioners and work in the field of teacher training/teacher education. As in the pilot study, the principal aim of the effectiveness study is to examine the implementation of the revised observation form and the adapted support system, covering issues like implementation fidelity. In addition, a second goal of the study is to examine the effectiveness of the workshop.

To assess the treatment effects on rater agreement, we will conduct a randomized control group design with repeated measures (see Gottfredson et al., 2015; Humphrey et al., 2016). Based on the current, ongoing demand for online teacher training and the parallel discussion concerning their effectiveness, we are going to compare the online workshop tested in the pilot study with a treated control group that receives a workshop targeting the same outcome variables but conceptualized as a traditional in-person block event. Research comparing the effectiveness of online and face-to-face training programs showed that both conceptualizations can be comparably effective when certain core components are taken into account in the conceptualization (see Fishman et al., 2013; Lipowsky & Rzejak, 2021). Thus, the effectiveness study will compare the effects of both workshop versions to derive implications for the further implementation process. Forty teacher trainers will participate in the study. To broaden the subject specificity, half of the participants will be teacher trainers in the field of math; the other half will be teacher trainers in the field of German.

As described, the project will follow the stepwise approach suggested for the conceptualization, implementation, and evaluation of interventions (Gottfredson et al., 2015; Humphrey et al., 2016). After the effectiveness study, a further evaluation study is planned to accompany the broad dissemination of the form, supplementary material, and workshop in the field (Step 6). The evaluation will focus on how the instrument is used in real-life classroom observations and how useful it is in providing teachers with effective feedback on teaching quality. This procedure will enable us to acquire knowledge reliably and repeatedly on whether the intended goals of the projects have been achieved and whether further adjustments are necessary.

12.5 Discussion

The "Promoting Teaching Quality through Classroom Observation and Feedback" project has three aims: First, as a rather specific goal, we hope to improve teaching quality among the teachers who receive feedback based on our observation form. The observation form can be used in different contexts: teacher education, teacher training, and peer feedback. Hence, the instrument was explicitly not developed to evaluate teacher performance in a high-stakes sense. Instead, it is a tool that supports further development and self-reflection and whose application is voluntary. Thus, teachers decide for themselves when, how often, and in which context they would like to receive feedback on their teaching in class. This approach was chosen to increase the acceptance of the form and its use in practice.

In addition to the desired effects on teaching quality, the project aims to increase students' performance. Previous studies have shown that the aspects of teaching quality considered in the form are predictive of student outcomes (e.g., Fauth et al., 2019; Lipowsky et al., 2009). Based on these study results, we assume an increase in students' performance in the long run and hope that the form will also be fruitful for discussions about which aspects of teaching quality should be considered in professional development programs and in everyday school practice. To verify this assumption, complex and long-lasting studies are necessary, including robust study designs. The research conducted within this project will reveal the extent to which these goals can be achieved.

As a third and a more general goal of the project, we hope to move the discussion about teaching quality within our state toward more research-based approaches. This aim is framed by the overall aim of the new educational quality policy established by the state government of Baden-Württemberg. That is, teaching practice should be based on the current state of science and empirically validated findings (Ministerium für Kultus, Jugend und Sport Baden-Württemberg, 2017). The development of a scientifically based instrument is a crucial step toward a common understanding of teaching quality. We hope that this observation form, the accompanying materials, and the workshops will provide opportunities for self-reflection and conversations about teaching quality. Drawing attention to these aspects could initiate a development in teaching practice that promotes teaching quality.

Until now, the project has involved a rather small and selected sample of teacher trainers/teacher educators in the field of mathematics. In the next step, the project team has now started to present the observation form and the accompanying material to a wider audience. We invited teachers, principals, teacher educators/teacher trainers, and school administrators to participate in information events, held in April/May 2021. The response to the project presentation was consistently positive and supported by a high level of interest. The effort to provide an observation form that can be used to examine teaching quality across subjects and school types was strongly welcomed. This positive feedback underscores the importance of the observation form and the project goals.

Acknowledgements We thank our colleagues at the Center for School Quality and Teacher Education (ZSL) for the excellent cooperation. We also thank the participants of the pilot and validation study for their interest and engagement in our project and all the fruitful discussions.

References

Andersson, C., & Palm, T. (2018). Reasons for teachers' successful development of a formative assessment practice through professional development—A motivation perspective. *Assessment in Education: Principles, Policy & Practice, 25*(6), 576–597. https://doi.org/10.1080/0969594X.2018.1430685

Bell, C. A., Qi, Y., Croft, A. J., Leusner, D., McCaffrey, D. F., Gitomer, D. H., & Pianta, R. C. (2014). Improving observational score quality: Challenges in observer thinking. In K. A. Kerr, R. C. Pianta, & T. J. Kane (Eds.), *Designing teacher evaluation systems: New guidance from the measures of effective teaching project* (1st ed., pp. 50–97). Jossey-Bass.

Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability, 21*(1), 5–31. https://doi.org/10.1007/s11092-008-9068-5

Briesch, A. M., Chafouleas, S. M., Neugebauer, S. R., & Riley-Tillman, T. C. (2013). Assessing influences on intervention implementation: Revision of the usage rating profile-intervention. *Journal of School Psychology, 51*(1), 81–96. https://doi.org/10.1016/j.jsp.2012.08.006

Brophy, J. (2000). *Teaching*. International Academy of Education.

Burke, M. J., Finkelstein, L. M., & Dusig, M. S. (1999). On average deviation indices for estimating interrater agreement. *Organizational Research Methods, 2*(1), 49–68. https://doi.org/10.1177/109442819921004

Darling-Hammond, L., Hyler, M. E., & Gardner, M. (2017). *Effective teacher professional development*. Learning Policy Institute. https://static1.squarespace.com/static/56b90cb101dbae64ff707585/t/5ade348e70a6ad624d417339/1524511888739/NO_LIF%7E1.PDF

Emmer, E. T., & Stough, L. M. (2001). Classroom management: A critical part of educational psychology, with implications for teacher education. *Educational Psychologist, 36*(2), 103–112. https://doi.org/10.1207/S15326985EP3602_5

Fauth, B., Decristan, J., Decker, A.-T., Büttner, G., Hardy, I., Klieme, E., & Kunter, M. (2019). The effects of teacher competence on student outcomes in elementary science education: The mediating role of teaching quality. *Teaching and Teacher Education, 86*, 102882. https://doi.org/10.1016/j.tate.2019.102882

Fauth, B., Decristan, J., Rieser, S., Klieme, E., & Büttner, G. (2014). Student ratings of teaching quality in primary school: Dimensions and prediction of student outcomes. *Learning and Instruction, 29*, 1–9. https://doi.org/10.1016/j.learninstruc.2013.07.001

Fishman, B., Konstantopoulos, S., Kubitskey, B. W., Vath, R., Park, G., Johnson, H., & Edelson, D. C. (2013). Comparing the impact of online and face-to-face professional development in the context of curriculum implementation. *Journal of Teacher Education, 64*(5), 426–438. https://doi.org/10.1177/0022487113494413

Gottfredson, D. C., Cook, T. D., Gardner, F. E. M., Gorman-Smith, D., Howe, G. W., Sandler, I. N., & Zafft, K. M. (2015). Standards of evidence for efficacy, effectiveness, and scale-up research in prevention science: Next generation. *Prevention Science: The Official Journal of the Society for Prevention Research, 16*(7), 893–926. https://doi.org/10.1007/s11121-015-0555-x

Hattie, J. (2010). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement* (Reprinted.). Routledge.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81–112. https://doi.org/10.3102/003465430298487

Humphrey, N., Lendrum, A., Ashworth, E., Frearson, K., Buck, R., & Kerr, K. (2016). *Implementation and process evaluation (IPE) for interventions in education settings: A synthesis of the literature.* Education Endowment Foundation. https://educationendowmentfoundation.org.uk/public/files/Evaluation/Setting_up_an_Evaluation/IPE_Review_Final.pdf

Kennedy, M. M. (2016). How does professional development improve teaching? *Review of Educational Research, 86*(4), 945–980. https://doi.org/10.3102/0034654315626800

Klieme, E., Pauli, C., & Reusser, K. (2009). The pythagoras study: Investigating effects of teaching and learning in Swiss and German mathematics classrooms. In T. Janik & T. Seidel (Eds.), *The power of video studies in investigating teaching and learning in the classroom* (pp. 137–160). Waxmann.

Klieme, E., Pauli, C., & Reusser, K. (2014). Unterrichtsbeobachtung - Pythagoras: Pythagorasmodul [Classroom observation - Pythagoras: Pythagorasmodule]. *DIPF German Institute for International Educational Research.* https://doi.org/10.7477/1:1:1

Klieme, E., Schümer, G., & Knoll, S. (2001). Mathematikunterricht in der Sekundarstufe I: „Aufgabenkultur" und Unterrichtsgestaltung [Mathematics teaching in lower secondary school: "Task culture" and lesson design]. In Bundesministerium für Bildung und Forschung (Ed.), *TIMSS – Impulse für Schule und Unterricht: Forschungsbefunde, Reforminitiativen, Praxisberichte und Video-Dokumente* (BMBF Publik, pp. 43–57).

Kounin, J. S. (1970). *Discipline and group management in classrooms.* Holt Rinehart & Winston.

Kraft, M. A., Blazar, D., & Hogan, D. (2018). The effect of teacher coaching on instruction and achievement: A meta-analysis of the causal evidence. *Review of Educational Research, 88*(4), 547–588. https://doi.org/10.3102/0034654318759268

Kunter, M., Klusmann, U., Baumert, J., Richter, D., Voss, T., & Hachfeld, A. (2013). Professional competence of teachers: Effects on instructional quality and student development. *Journal of Educational Psychology, 105*(3), 805–820. https://doi.org/10.1037/a0032583

Kyriakides, L., Christoforou, C., & Charalambous, C. Y. (2013). What matters for student learning outcomes: A meta-analysis of studies exploring factors of effective teaching. *Teaching and Teacher Education, 36*(1), 143–152. https://doi.org/10.1016/j.tate.2013.07.010

Lipowsky, F., Rakoczy, K., Pauli, C., Drollinger-Vetter, B., Klieme, E., & Reusser, K. (2009). Quality of geometry instruction and its short-term impact on students' understanding of the Pythagorean Theorem. *Learning and Instruction, 19*(6), 527–537. https://doi.org/10.1016/j.learninstruc.2008.11.001

Lipowsky, F., & Rzejak, D. (2021). *Fortbildungen für Lehrpersonen wirksam gestalten* [Effectively design teacher trainings]. Bertelsmann Stiftung 2021. https://doi.org/10.11586/2020080

Lotz, M., Lipowsky, F., & Faust, G. (Eds.). (2013). *Materialien zur Bildungsforschung: Vol. 23, 3. Dokumentation der Erhebungsinstrumente des Projekts „Persönlichkeits- und Lernentwicklung von Grundschülern" (PERLE): 3. Technischer Bericht zu den PERLE-Videostudien [Documentation of the survey instruments of the project "Personality and Learning Development of Elementary School Students" (PERLE)].* Gesellschaft zur Förderung Pädagogischer Forschung [u.a.].

Maaß, K., & Artigue, M. (2013). Implementation of inquiry-based learning in day-to-day teaching: A synthesis. *ZDM Mathematics Education, 45*(6), 779–795. https://doi.org/10.1007/s11858-013-0528-0

Ministerium für Kultus, Jugend und Sport Baden-Württemberg. (2017, June 28). *Neues Qualitätskonzept für das Schulsystem [New quality concept for the school system]* [Press release]. Stuttgart. https://www.baden-wuerttemberg.de/de/service/presse/pressemitteilung/pid/qualitaetskonzept-fuer-das-bildungssystem-baden-wuerttembergs/

OECD. (2013). Teachers for the 21st Century: Using evaluation to improve teaching. *OECD*. https://doi.org/10.1787/9789264193864-en

Parchmann, I., Gräsel, C., Baer, A., Nentwig, P., Demuth, R., & Ralle, B. (2006). "Chemie im Kontext": A symbiotic implementation of a context-based teaching and learning approach. *International Journal of Science Education, 28*(9), 1041–1062. https://doi.org/10.1080/09500690600702512

Pianta, R. C., & Hamre, B. K. (2009). Conceptualization, measurement, and improvement of classroom processes: Standardized observation can leverage capacity. *Educational Researcher, 38*(2), 109–119. https://doi.org/10.3102/0013189X09332374

Praetorius, A.-K., Klieme, E., Herbert, B., & Pinger, P. (2018). Generic dimensions of teaching quality: The German framework of three basic dimensions. *ZDM Mathematics Education, 50*(3), 407–426. https://doi.org/10.1007/s11858-018-0918-4

Rakoczy, K., & Pauli, C. (2006). Hoch inferentes Rating: Beurteilung der Qualität unterrichtlicher Prozesse [High-inference rating: Assessing the quality of instructional processes]. In I. Hugener, C. Pauli, & K. Reusser (Eds.), *Materialien zur Bildungsforschung: Vol. 15. Dokumentation der Erhebungs- und Auswertungsinstrumente zur schweizerisch-deutschen Videostudie „Unterrichtsqualität, Lernverhalten und mathematisches Verständnis". Teil 3. Videoanalysen* (pp. 206–233). GFPF.

Seidel, T., & Shavelson, R. J. (2007). Teaching effectiveness research in the past decade: The role of theory and research design in disentangling meta-analysis results. *Review of Educational Research, 77*(4), 454–499. https://doi.org/10.3102/0034654307310317

Stanat, P., Schipolowski, S., Rjosk, C., Weirich, S., & Haag, N. (Eds.) (2017). *IQB-Bildungstrend 2016: Kompetenzen in den Fächern Deutsch und Mathematik am Ende der 4. Jahrgangsstufe im* *zweiten Ländervergleich [IQB-Bildungstrend 2016: Competences in German and mathematics at the end of grade 4; second country comparison].* Waxmann.

Statistisches Landesamt Baden-Württemberg. (2020a). *Allgemeinbildende Schulen: Allgemeinbildende Schulen nach Schularten [Schools of general education according to type of school].* https://www.statistik-bw.de/BildungKultur/SchulenAllgem/abschulen.jsp

Statistisches Landesamt Baden-Württemberg. (2020b). *Allgemeinbildende Schulen: Lehrkräfte nach Beschäftigungsverhältnis [Schools of general education: Teachers by employment status].* https://www.statistik-bw.de/BildungKultur/SchulenAllgem/ablehrer.jsp

Taut, S., & Rakoczy, K. (2016). Observing instructional quality in the context of school evaluation. *Learning and Instruction, 46*, 45–60. https://doi.org/10.1016/j.learninstruc.2016.08.003

Wisniewski, B., & Zierer, K. (2020). Entwicklung eines Online-Fragebogens zur Erhebung von Unterrichtsqualität durch Lernendenfeedback und erste Validierungsschritte [Development of an online questionnaire to survey teaching quality through learner feedback and first validation steps]. *Psychologie in Erziehung Und Unterricht, 67*, 138–155. https://doi.org/10.2378/peu2020.art10d

Part V
Teacher Evaluation Systems Around the World: Asia and Oceania

Chapter 13
An Overview of the Teacher Evaluation System in China

Gang Li and Tao Xin

Abstract This chapter outlines and appraises the teacher evaluation system in China. Firstly, the historical review finds that the evolution of the system since the late 1970s can be divided into three stages. We present several major areas pertaining to the current system including evaluation of teacher qualifications, performance evaluation, teachers' professional title evaluation, appraisal for awards and work excellence titles, and classroom teaching evaluation. A detailed description of the evaluation contents and indicators, procedures and methods, and key actors is provided. The chapter, substantiated by related empirical studies, analyzes the effects of the teacher evaluation system on school management and teachers' professional development, the shortages in terms of the functions, contents and indicators, methods and instruments, usage of the evaluation results, and burdens caused by evaluation. Based on the review, this chapter suggests that China should reform its current teacher evaluation system in the new era and demonstrates some of the future reform directions.

13.1 Introduction

Teacher evaluation refers to the process of judging teachers' actual work and its potential value, which is not only closely related to teachers' salaries, promotions, and other interests, but also a strong incentive to their professional development (e.g., Organisation for Economic Co-operation and Development (OECD), 2013, p. 41; Stronge, 2006; Wise et al., 1985). Globally, teacher evaluation has gained attention and generated heated debate. How to establish a fair and scientific teacher evaluation system has been a huge challenge to both policymakers and school authorities (Liu & Zhao, 2013). Before the 1980s, there were few regular approaches to evaluating elementary and middle school teachers in the People's Republic of China (hereafter

G. Li (✉) · T. Xin
Beijing Normal University, Beijing, China
e-mail: ligang@bnu.edu.cn

T. Xin
e-mail: xintao@bnu.edu.cn

J. Manzi et al. (eds.), *Teacher Evaluation Around the World*,
Teacher Education, Learning Innovation and Accountability,
https://doi.org/10.1007/978-3-031-13639-9_13

"China"). However, since the profound reform and opening up since 1978,[1] a teacher evaluation system within the scope of basic education has been gradually established and refined. By reviewing the evolution of teacher evaluation system in China as well as its effectiveness and remaining deficiencies, this chapter attempts to depict a comprehensive and vivid picture of China's reform on teacher evaluation.

13.2 Background of China's Teacher Evaluation

China has the world's largest population of school-aged children. According to the statistics of China's Ministry of Education, there were around 530,100 institutes of all kinds at all levels, with 17,320,300 teachers, providing education services to over 282 million students in 2019 (Ministry of Education, 2020a).

13.2.1 Education System

China's education system has implemented nine-year compulsory education for all school-aged children since 2006. Students need to complete six years of elementary education (ISCED 1) and three years of lower secondary education (ISCED 2). At this educational stage, the Chinese government adopts the nearby enrollment policy, which allocates students to the nearest schools instead of letting them take screening tests. However, many private and public lower secondary schools, which are so-called good schools, still organize screening tests. After completing lower secondary schooling, students need to take upper secondary school entrance examination (*Zhongkao*) in order to participate in upper secondary education (ISCED 3). The upper secondary education includes two major learning tracks: general education programs and vocational education programs. In general, high performers tend to enter general upper secondary schools. After three years of schooling in upper secondary school, students sit for national university entrance examination (*Gaokao*) to compete for higher education opportunities. In China's education system, these tests not only have a huge impact on students' academic career and future life, but also influence the evaluation of education quality at schools.

Public education has a dominant share in China's education system (see Table 13.1). In contrast to public schools, private schools are in the minority, but enjoy greater autonomy in curriculum, teaching, and administration. However, according to the *Non-state Education Promotion Law*, "Non-state education is a public undertaking. It's a part of the socialist education undertakings. The country adopts the policies of active encouragement, full support, correct guidance, and administration by law" (Standing Committee of the National People's Congress, 2018), and thus government policies still have tremendous guidance on private schools. Governments at various levels have also intensified control over private schools through grants and procurement of services in the recent decade.

Table 13.1 Number and share of China's public and private schools in basic education in 2019

	Run by government		Run by non-government		Total
	Number	Percent (%)	Number	Percent (%)	Number
Elementary schools	250,271	97.5	6333	2.5	256,604
Junior secondary	46,622	88.9	5793	11.1	52,415
Senior secondary	10,522	75.4	3427	24.6	13,949
Secondary vocational schools	5610	73.9	1985	26.1	7595

Note Adapted from Ministry of Education, http://www.moe.gov.cn/s78/A03/moe_560/jytjsj_2019/qg/

13.2.2 School Teachers

Teachers are the bedrock that underpin and sustain China's huge education system. As of 2019, there were around 6,269,100 elementary school teachers, 3,747,400 lower secondary school teachers, and 1,859,200 upper secondary teachers. Among all, 99.97% of elementary school teachers held a teacher qualification certificate, and the percentages for lower secondary school teachers and upper secondary school teachers were 99.88% and 98.62%, respectively. The number of secondary vocational school teachers reached 842,900 in the same year (Ministry of Education, 2020a, 2020b).

China established a multi-approach teacher education system to cultivate teachers. College students who want to be a teacher can either receive teacher education or take extra courses including education, psychology, or related fields and pass the evaluation of teacher qualifications. Those who already have a job are eligible for the application of a teacher certificate as well. When successfully becoming a teacher at school, one receives induction and mentoring support from the school to adapt to the role of a teacher. Throughout their careers, teachers have access to regular professional development training and collaborative learning opportunities that help them develop the skills needed to overcome new challenges (OECD, 2020, p. 18).

13.2.3 Education Governance

At first sight, China's education system adopts a top-down approach, which means that the lower levels of local government and schools seemingly just implement what the higher levels have decided. However, the fact is that the governance over basic education is more complex than what it appears to be. It is true that governments have significant influences on schools, but their roles vary. The central government sets out the strategic directions for education reform, and local governments need to implement central government's policies by taking account of the local conditions of social, economic, and educational development. In China, governments at the county level are in charge of compulsory education, while those at the municipal level take charge of upper secondary schooling. Collectively, they formulate implementation opinions to make sure that policies issued by the higher governments can be enacted at schools. Furthermore, they are responsible for specific managerial tasks such as appropriation, selection and appointment of headmasters, and supervision.

Over the past three decades, the educational system in China has undergone a decentralization trajectory. Schools are granted a great deal of autonomy in making decisions on teaching, personnel management, and the use of funds (Ministry of Education et al., 2020). Teachers not only have autonomy in teaching and professional development, but also can participate in school management as teacher leaders or through teacher representative assemblies and all-teacher meetings. In reality, due to some governments' over-regulating and rigid management style, it is difficult to guarantee the autonomy of school running, which demotivate relevant schools for independent development and innovation but precipitate these schools to habitually obey governments' arrangements in all aspects (Chu, 2008). Sometimes teachers are also unwilling to speak up on school affairs due to heavy workload, tradition of respect for authority, and lack of skills to participate in management (Carney, 2009; Lai & Lo, 2007).

13.3 Evolution of the Teacher Evaluation System in China

According to the major policies implemented in China related to teacher evaluation, the development of China's teacher evaluation system since the late 1970s can be divided into three periods, as summarized in Table 13.2.

Table 13.2 Three periods of China's teacher evaluation system development

Period	Reform background	Related evaluation systems
Early stage of China's teacher evaluation system (Late 1970s–2001)	Administration of teaching force needs to be enhanced after the cultural revolution	Initial form of teacher appraisal Evaluation of teacher qualifications Teachers' professional title evaluation
Teacher evaluation in the context of the new curriculum reform (2001–2009)	As a new round of curriculum reform begins, teachers' professional development needs to be promoted	Periodic teacher appraisal
Teacher evaluation along with the performance-related pay reform (since 2009)	The government implemented performance-related pay system in health, culture, and education sectors. The system in the education sector aims to raise teachers' salaries and stimulate their vitality	Performance-related teacher evaluation Classroom teaching evaluation

13.3.1 Early Stage of China's Teacher Evaluation System (Late 1970s–2001)

From 1966 to 1976, China experienced a turbulent decade of the Cultural Revolution, causing disastrous damage to the country's education system. At the end of 1978 when China began to carry out the reform and opening up, it was an imperative to put education back on track, especially to strengthen the management of the teacher workforce. According to *China Education Statistical Yearbook*, only 47.0% of primary school teachers had a high school degree or higher, and 7.9% of junior high school teachers and 50.8% of senior high school teachers had an associate degree or higher (Department of Planning, Ministry of Education, 1980, pp. 86–87).Teacher evaluation in this period served as an effective tool to enhance teacher management, with measures taken mainly from two facets.

The first facet was to enhance the overall quality of the teaching force by teacher appraisal. According to the *Recommendations on Strengthening Administration of Teachers in Elementary and Secondary Schools* issued by the State Education Commission (Adjusted and renamed to the Ministry of Education in 1998) in 1983, educational administrative departments at the county level should evaluate teachers from morality and working attitude, teaching skills and outcome, as well as educational level. And the results are taken as references for providing teacher training and making decisions about teacher recruitment and position adjustment (State Education Commission, 1983). This document for the first time emphasized the necessity of regular teacher evaluation. In 1985, the *Seminal Decision on Reform of Educational System* demands to evaluate all in-service teachers thoroughly and offer training to them in the next five years or longer (Central Committee of Chinese Communist Party (CPC), 1985). Although relevant requirements on teacher evaluation were put forward, governments at the county level were unable to formulate specific evaluation criteria or develop professional evaluation instruments. In practice, they concentrated mainly on teachers' educational level, an indicator that is easy to evaluate. Based on the appraisal results, the local governments provided professional development training for teachers accordingly, especially with a focus on upgrading their educational level, or eliminating underperforming teachers to make sure that all teachers at their posts meet the baseline requirements of being teachers.

The second facet was to establish the teacher qualification and professional title system. In 1986, the *Compulsory Education Law* states that all professional teachers shall obtain the state-regulated teacher qualification certificate (National People's Congress, 1986). In the same year, the SEC issued the *Interim Measures for Qualification Certificates for Elementary and Secondary School Teachers*, marking the nationwide implementation of the teacher qualification system. In 1993, the *Teacher Act* articulates that the state shall institute a system of qualifications for teachers, which gave the legal recognition of the system (Standing Committee of the National People's Congress, 1993). The *Regulations on the Qualifications of Teachers* issued in 1995 further clarify the qualifications for teachers in schools of all kinds and levels as well as the corresponding procedures of the teacher qualification examination and certification (State Council, 1995). Accordingly, prospective teachers, in addition to meeting certain requirements, must pass the teachers' qualification tests and the assessment of trial lecture. While the education department of the State Council determines and approves the test subjects, standards, and syllabus of the tests and trial lecture, municipal or provincial education departments are responsible for putting them into practice. *Interim Rules for the Duties of elementary School Teachers* and *Interim Rules for the Duties of Middle School Teachers,* respectively, issued in 1986 define the required expertise and qualifications for teachers in different stages of professional development, which marked the formal establishment of the teachers' professional title system (State Education Commission, 1986a, 1986b). As per those requirements, teachers must prove their professionalism by submitting a series of documents, which will be reviewed by evaluation committee for professional title at school and municipal government levels. A debriefing ensues.

During this stage, as policies regarding teacher appraisal, teacher qualification, and teachers' professional title were enacted, teacher evaluation system was gradually established.

13.3.2 Teacher Evaluation in the Context of the New Curriculum Reform (2001–2009)

In 2001, the Ministry of Education (MOE) issued the *Guideline on the Reform of Curriculum in Basic Education*, and the new round of curriculum reform in basic education was launched. The reform targeted at converting China's traditional test-oriented education system into a quality education system where knowledge-based teaching, passive learning, and "teach to the test" were abandoned; instead, the philosophy of nurturing all-around students and preparing them for their future life was adopted (Huang, 2004; Liu & Teddlie, 2003). As one of the most radical and complex education reforms in the world, it has not only brought new evaluation ideas, but also advanced new curriculum reform by cultivating more professional teachers facilitated by teacher evaluation. The *guideline* underlines:

> To establish an evaluation system that can constantly enhance teacher's competence, promote teachers' reflection and analysis on their own teaching practice; and to build an evaluation system based on teachers' self-evaluation along with the involvement of various stakeholders including principals, parents and students. Thus, teachers could improve their expertise prompted by information and feedback from multiple channels (Ministry of Education, 2001).

In 2002, the MOE issued the *Notification on Advancing Reform of Evaluation and Exam System for Elementary and Secondary Schools*, claiming to establish an evaluation system that is conducive to building on teachers' professional morality and competence. Schools should evaluate teachers by considering their professional ethics, the extent to which they understand and respect students, the competence of designing and carrying out teaching plans, and the effectiveness of communication and reflections. Such evaluation system prioritizes both teachers' self-assessment and the feedback from school administrators, colleagues, students, and parents. Furthermore, students' test performance is not allowed to be used as the sole criteria to assess teachers (Ministry of Education, 2002). In 2003, the MOE issued the *Notice on Further Strengthening the Management and Professional Morality Education of Elementary and Secondary School Teachers*, underscoring again the necessity of improving teacher evaluation system to be co-participated by school leaders, teachers, parents, and students on the basis of self-evaluation (Ministry of Education, 2003).

Under such policy background, schools in China assessed teachers' performance on a monthly or yearly basis with continuous self-driven improvement. Such appraisal focused on the following four aspects: (1) morality—teachers were required to abide by the law and disciplines and stick to professional ethics; (2) ability—teachers were expected to be competent in classroom teaching, pedagogical research, and classroom management; (3) diligence—teachers were required to attend lessons on time and get involved in various activities organized by the schools; (4) achievements—teachers were evaluated for their achievements from various aspects of their daily works, including students' scores in *Zhongkao*, *Gaokao,* and other mid-term and final term tests, teachers' own awards and honorable titles, and publications. Based on this general evaluation framework, school authorities further developed more detailed indicators. However, some of the indicators were very abstract and difficult to operationalize (Liu & Teddlie, 2005).

In this stage, the central government introduced policies that provided guidelines about evaluators and evaluation content of teacher evaluation, but different schools and regions carried them out differently. In some regions, teachers' performances were still mainly rated by school administrative team in a simple form. But in relatively developed regions like Beijing, Shanghai, Zhejiang, and Jiangsu, local governments and schools carried out profound reform on teacher evaluation, particularly in evaluation organization and approach (e.g., Zhang & Ng, 2011). Besides from school administrative teams, teachers themselves, colleagues, parents, and students were involved in the evaluation. Teachers had to submit a self-reflection report, or create a professional portfolio, which includes records of their trainings, teaching researches, papers, and honors. Colleagues especially head teachers of the teaching and research groups need to rate teachers' performance. By completing simple designed questionnaires, students could rate classroom teaching, and parents could express their levels of satisfaction with teachers. At the later stage of this period, classroom observation became a popular evaluation method. As Zhang and Ng (2011) commented on the effect of Shanghai's teacher evaluation system in this period, the system has created pressure and extrinsic incentives for teachers to improve and provided them with guidance and directions. However, in many cases, due to the lack of standardized indicator system and evaluation instruments, this type of evaluation did not have much impact. It was even possible that high scores would be given in turns to different teachers within a department, resulting in the tokenism of evaluation.

13.3.3 Teacher Evaluation Along with the Performance-Related Pay Reform (since 2009)

When it comes to teacher evaluation, performance-related pay has been a trending issue in European and American countries, but it was put on agenda and investigated relatively late in China. In 2008, the central government decided to implement performance-related pay in health, culture, and education sectors. In December 2008, the *Guiding Opinions on Implementing Performance-Based Pay in Compulsory Education Schools* (State Council, 2008) and the *Guiding Opinions of the Ministry of Education on Conducting Teacher Performance evaluation in Compulsory Education Schools* (Ministry of Education, 2008) were issued, mandating that the performance-related pay system and performance appraisal shall be applied from 2009, and since then it has become the mainstream in China's teacher evaluation system. Under the performance-related pay system, teachers' salaries are divided into base pay (70%) and merit pay (30%), and the latter is allocated according to the results of performance evaluation (State Council, 2018). By setting up the performance-related pay system, the central government strived to avert the tendency of "giving priority to seniority" and "equalitarianism" in salary allocation, so as to raise teachers' salaries and encourage them to pursue better performance.

Against such a backdrop, teacher evaluation system has undergone two major changes. On the one hand, teacher evaluation paid more attention to teachers' actual performance in completing their duties and responsibilities, including morality, accomplishment in teaching or serving as a class head teacher, which embodied an apparent favor of result-oriented and quantitative approach. On the other hand, evaluation results were directly linked to teachers' salaries and became a strong justification for recruitment, training, and promotion, thus it became a high-stakes evaluation. Performance evaluation has to some extent raised the salaries of some proportion of teachers in localities. The move further narrowed the gap between the average salaries of teachers in compulsory education schools and the salaries of civil servants in the same region, which to an extent boosted teacher morale. However, within schools, problems still exist in regard to the distribution of the merit pay for teachers. In 2020, the Central Committee of the CPC and the State Council issued the *Overall Plan for Deepening Educational Evaluation Reform in the New Era*, which stresses the urgency of improving the performance evaluation. Nonetheless, specific measures remain to be enacted.

13.4 A Glance at China's Current Teacher Evaluation System

Teacher evaluation should run through teachers' entire careers and professional development (Darling-Hammond, 2012). China has now established a teacher evaluation system throughout the entire teacher career, from determining to be a teacher to pursuing constant growth and professional development. In China's current teacher evaluation system, there are five most common categories of teacher evaluation. All the teachers of basic education must get through teacher qualification evaluation when they want to be a teacher and get involved in performance evaluation and teachers' professional title evaluation once a year. Outstanding teachers voluntarily accept appraisals for awards and work excellence titles several times a year. Classroom teaching evaluation varies among different schools.

13.4.1 Evaluation of Teacher Qualifications

Teacher qualification evaluation is an official evaluation of applicants to assess whether they are qualified to be teachers. According to the *Regulations on the Qualifications of Teachers* (State Council, 1995), teacher qualification evaluation mainly focuses on the following three aspects: (1) basic capability which is essential to education and teaching. Based on those guidelines, the Ministry of Education develops the *Standards for the Teacher Qualification Examination* (Department of Teacher Education, Ministry of Education & National Education Examinations Authority, 2011) and outlines for tests and trial lecture (Department of Teacher Education, Ministry of Education & National Education Examinations Authority, 2012a, 2012b). The outline for tests specifies the content, proportion of different parts, and item types, while the outline for trial lecture regulates the content, methods, and scoring rubrics. Specific tests are formulated by the provincial education administrative department. However, there is a rising number of provinces using the national teacher qualification certificate tests. Applicants must take two subtests: comprehensive quality and knowledge and skills in education. Applicants for lower and upper secondary teacher certificate need to take an additional subtest: subject-related knowledge and teaching ability. A face-to-face trial lecture follows to assess applicants' practical abilities. Municipal administrative departments for education select a panel of professors, K-12 school teachers, and educational research experts who must be certified by receiving training from examination institutions at the provincial level or above. The trial lecture, which includes teaching and structured defense, will be graded in occupational understanding, psychological quality, deportment, verbal expressions, traits of thinking, teaching design, teaching practice, teaching evaluation, etc. (2) Proficiency in Mandarin—applicants must have obtained Level 2(B) or above, proving their ability to deliver teaching in standard Mandarin without strong accents. Yet teachers from ethnic minority areas may have such requirement loosened. (3) Physical and mental fitness—applicants must receive medical examination in qualified hospitals, proving themselves free from serious or infectious diseases or disabilities that hamper teaching, such as stammer, visual impairment, and hearing disorder. Local governments will check if applicants suffer from mental diseases during qualification examination. (4) Educational level—according to the *Teacher Act* (Standing Committee of the National People's Congress, 1993), to be qualified for a teacher in elementary school, one shall be a graduate of a secondary normal college or higher; to be qualified for a teacher in junior middle school, one shall have an associate degree from a higher normal or other university and higher; to be qualified for a teacher in senior high school, one shall get a bachelor's degree from a higher normal or other university. In 2013, the MOE issued the *Interim Measures for Regular Registration of Teacher Qualifications in Elementary and Secondary Schools*, saying that local governments should renew teachers' qualification certificates every five years with the focus on the two following aspects: First, they do not violate teacher ethics. Second, they are required to get mandatory training within five years (Ministry of Education, 2013).

The establishment of the teacher qualification evaluation system has clarified the basic requirements for being a teacher and enhanced the overall competence of the teaching body in China. However, in terms of educational level of the employed teachers, there seems to be a higher standard in many other countries. In 2018, the average proportion of junior high school teachers with a master's degree is 44.2% in OECD countries, 54.9% in EU countries, over 60% in such developing countries as Bulgaria and Latvia, and over 20% in Romania and Mexico. In contrast, in 2019, only 3.5% of teachers working at junior high schools held a master's degree in China, and the proportion for senior high school teachers was 10.6% (Wang, 2021). Consequently, there has been a growing call for raising the requirement of teachers' degree levels, and some propose that the minimum threshold for being an elementary or middle school teacher should be holding a master's degree.

13.4.2 Performance Evaluation

Performance-based pay system and performance evaluation have been adopted since 2009, which gradually incorporated the previous periodic appraisal in most schools. The *Guiding Opinions* (Ministry of Education, 2008) on conducting teacher performance evaluation are premised on the concept that the evaluation shall take various aspects including teacher ethics, education and teaching work, teaching effectiveness, and the professionalism of being a class teacher into account. The evaluation of teacher ethics is based on whether elementary and secondary school teachers follow the required professional ethics, which is also a prerequisite for every educator to pass the performance evaluation. Education and teaching focus on teachers' performance in moral education, teaching strategies, pedagogical research, and professional development. The evaluation of teachers' effectiveness focuses on whether teachers complete the targets set by the state and whether their students' school performance can meet the prescribed minimum standards, but the proportion of students entering high schools or universities (promotion rate) shall be excluded from the evaluation indicators. The professionalism of being a class teacher focuses on teachers' education and guidance for students, class management, the organization of class activities, and other forms of collective activities. It should be noted that the *Guiding Opinions* simply listed key focuses of performance evaluation without specifying indicators and methods. Governments at the county level introduced implementation opinions, with some specifying evaluation indicators and methods for schools to put into practice, whereas schools set up leading groups responsible for formulating detailed rules to carry out the evaluation. Procedurally, the performance evaluation plan needs to be deliberated and approved by the faculty representative assembly or the faculty assembly. Teachers are extensively involved in the process of formulating the plan for performance evaluation. Some principals interviewed by the author said that "a performance evaluation plan is often discussed for many rounds by the faculty representative assembly and can be even overturned many times" (Li, 2016, p. 113).

According to some empirical investigations (Liu et al., 2016; Zhao et al., 2011), in practice, schools tend to value the following evaluation indicators in stronger terms: workload (credit hours), teaching performance (especially students' test scores and promotion rate), teacher ethics, work attitude, attendance, working as a class teacher, length of service, academic qualifications, publications and research projects, and parents' evaluation on teachers' performance. Among all these indicators, the top four prioritized indicators are workload, teaching quality, teacher ethics, and attendance (Zhao et al., 2011), which, to some degree, are consistent with the guiding opinions of the MOE. In some countries/regions, teachers' professional standards are a major basis for determining the content and specific indicators of performance evaluation (e.g., Evans, 2013; Taut et al., 2011). But in China, such standards were not available until 2012. Besides, since professional visions and values, skills, and knowledge are difficult to measure, they have not exerted a substantial impact on the performance evaluation. When it comes to the implementation of evaluation, the work is mostly undertaken by principals, members of the school leadership, and head teachers of the teaching and research groups. Other ordinary teachers, students, and parents seldom participate in the evaluation process. With respect to specific evaluation methods, the most common one is rating after reviewing students' test results, applicants' workload, awards, and papers.

13.4.3 Teachers' Professional Title Evaluation

Professional title is designed to reflect the degree of professionalism of those working in a given field. Teachers' professional title evaluation plays a prominent role in reflecting skills required for the job, guiding teachers' professional development and also influencing teachers' benefits. Currently in China, there is a hierarchy of professional titles for teachers which consists of five levels, with "senior professional" (equivalent to professorship) title topping the hierarchy, followed by "senior teacher", "first-grade teacher", "second-grade teacher", and "third-grade" teachers ranking from the highest professional level to the lowest.

In terms of the evaluation criteria, the *Basic Standards and Requirements for Teachers' Professional Title Evaluation in Elementary and Secondary Schools* clarify rules for the evaluation of teachers' professional titles at all levels, taking into account requirements including educational level, length of service as a teacher, teaching ability, proficiency in educational theories, skills on teaching research, and teaching effectiveness (Ministry of Human Resources and Social Security & Ministry of Education, 2015). Even though the *Basic Standards and Requirements* expressly state teachers' educational level, length of service, there are no mandatory rules on the last three ones. On this basis, governments at the provincial level should further prescribe the requirements. For instance, *Requirements of Applying for Teacher Professional Title in Elementary and Secondary Schools in Beijing* (Department of Human Resources & Social Security of Beijing, 2016) state a specific indicator for first-grade teachers' teaching ability, "Having won at least third-class prizes in teaching contest at district levels and above, or taught district-level open classes or demonstration lessons" (Department of Human Resources & Social Security of Beijing, 2016), (First-grade teacher, para. 5). In fact, due to a lack of tools that evaluate teaching ability and distrust of the tools, when provincial governments select specific indicators, great emphasis is put on more visible ones such as the years of teaching experience and the awards the applying teachers have won in various teaching contests. More often than not, teachers with more years of teaching experience are given priority consideration for title advancement when too many candidates compete for limited number of professional titles. Although professional titles should reflect the levels of teachers' professional development, current evaluation standards are not closely linked with the professional standards.

Municipal governments and schools take charge of the specific evaluation process. Schools are responsible for the evaluation of second-grade and third-grade title applicants. An evaluation panel is created, which includes school administrators, head teachers of teaching and research groups, and teachers. The panel reviews the documents submitted by applicants, taking into account their performance. Since the competition is not fierce, candidates meeting the minimum requirements can obtain the titles. But for applicants of the first-grade title or above, teachers are firstly nominated by the school authorities through a series of procedures, including filing an application form, reporting their work experience to the evaluation panel, going through a democratic assessment process, and finally obtaining evaluation panel's recommendation (Zhang, 2011). Municipal governments need to create a city-level evaluation committee consisting of officials from the municipal education department, teachers, and principals with seniority, etc. In addition to reviewing application forms, the panel also listens to applicants' "*Shuoke*" (orally presenting teaching plans and underlying understanding and values.) and lets them answer questions. Accordingly, the best applicants will be given first-grade teacher titles or above.

13.4.4 Appraisal for Awards and Work Excellence Titles

Appraisal for awards and work excellence titles is not necessarily implemented periodically, but it does have a close bearing on teachers' honors and salaries, as well as a strong sense of academic and moral achievement due to their exemplary effects, hence a strong incentive for applicants and awardees. At provincial levels, such honorable titles as “special-grade teachers”, “model teachers”, “advanced individuals”, “excellent teachers in the educational system”, and “model educators” are prominent awards and titles. These title owners are eligible to compete for national honorable titles. At the district and county levels, teachers place a higher value on awards such as “subject leading leaders” and “backbone teachers”. A subject leading teacher, referring to a teacher with expertise in a specific academic area, takes a keen interest in and has strong competence to conduct pedagogical research, achieve markable research results, and serve as role models to lead and organize teachers to improve their teaching and research. A backbone teacher refers to a teacher with abundant teaching experience and is exemplary for young teachers to learn from. Besides, those who display a high degree of morality and dedication to teaching career, as well as those newly recruited teachers who demonstrate excellent teaching performance are also commended with relevant honorable titles.

Two dimensions of requirements and indicators are adopted in the appraisal for awards and work excellence titles. Firstly, the candidate teachers have to meet some qualification-related prerequisites such as teachers’ professional titles and the years of service; secondly, the teaching experience, mainly manifested by the applicants’ previously earned awards in various teaching contests, plays a significant role in the application for the following awards. However, in terms of the appraisal and selection of the special-grade teacher titles, candidates need to develop their own educational philosophy, explore effective teaching methods, and achieve outstanding results in teaching, which have created a huge social impact. They need to submit supportive documents, such as papers, works, news coverages of their teaching practice, or even evidences showing that other schools and teachers have followed suit.

As for the procedures, local education authorities take charge of setting quotas in advance for how many applicants a school can recommend in proportion to the teacher population of the school. After reviewing the documents from candidates, principals, members of the school leadership, and head teachers of the teaching and research groups will decide nominees. In many schools, the principal’s opinions play a decisive role. Local governments choose the best ones to compete for higher-grade honorary titles, but will strike a balance between different types of schools. For instance, more quotas will be allocated to rural schools or disadvantaged urban schools.

13.4.5 Classroom Teaching Evaluation

Unlike the above-mentioned evaluation models, of which the results are closely linked to teachers’ promotions, salaries, and awards, the evaluation of classroom teaching is more related to teachers’ professional development. Chinese education always embraces the tradition of research and collaboration in teaching. Classroom observations, or classroom visits, are often used by examiners to conduct classroom teaching evaluation (OECD, 2020, pp. 130–131). This evaluation model features a strong focus on teachers' gradual development in professional terms prevails nationwide, especially after China's national curriculum reform.

Governments do not issue guidelines on the content and methods of classroom observation. Rather, they are up to schools and teachers. In terms of evaluation contents, while traditional classroom observations mainly focus on how teachers impart knowledge in the textbook, such as whether they can articulate specific knowledge points to students, greater attention has been given to how teachers motivate students to learn through inquiry and cooperation, as well as how they cultivate students' ability of independent thinking in the current evaluation system. A team of researchers jointly with elementary and secondary school teachers proposed the "LICC" classroom observations paradigm, which becomes a guideline for classroom observation and is promoted on a large scale in schools in Zhejiang and Shanghai (Cui et al., 2013; Shen & Cui, 2008). Under this paradigm, teachers need to pay attention to four elements (20 sub-elements): (1) learning: preparation before class, listening, interaction, self-directed learning, goal attainment; (2) instruction: teaching design, presentation, dialogue, student guidance, and teaching tact; (3) curriculum nature: goals, content, implementation, evaluation, and resource; (4) classroom culture: facilitating thinking, democracy, encouraging innovation, caring about students, and uniqueness of teaching and learning (Cui, 2012). Teachers only need to choose parts of the elements, catering for the evaluation purpose and individual preference. Table 13.3 shows an instrument for classroom observation created by a secondary school teacher and that is widely used in some schools. The observer will record the target of the teacher's gaze and its frequency at regular intervals. Teachers can better understand who or what they habitually pay attention to and whether they have neglected eye contact with student through the analysis of the record.

When put into practice, classroom observations are sometimes organized at the school level, requiring all teachers to take turns to sit in on another teacher's class for observation, learning, reflection, and filling certain observation reports. In some

Table 13.3 Instrument for classroom observation created by a secondary school teacher

What is the teacher looking at?	Frequency	Proportion
All the students		
Students in the front of the classroom		
Students in the middle of the classroom		
Students at the back of the classroom		
Students answering questions		
Students who are demonstrating		
Distracted or sleepy students		
The blackboard, computer, textbook, projector screen, etc.		
The ceiling or other things that have nothing to do with teaching		

Note Adapted from "Classroom observation II: Towards professional 'Tingpingke'", by Cui et al. (2013)

schools, teachers voluntarily organize classroom observation as means for mutual and collaborative learning. To carry out the evaluation, the observer teachers need to hold a meeting before the class visit to do certain preparation work, such as clarifying the key indicators and methods, and also dividing responsibilities to different observers; in the process of classroom observation, the observers need to pay special attention to evaluation indicators; after the classroom observation, a review meeting is normally held for the exchange of ideas, during which the observed teachers reflect upon and analyze their own strengths and weaknesses demonstrated in the teaching process, while the observers are required to give feedbacks based on the ratings for the classroom observation (Cui, 2012).

13.5 Effects of China's Teacher Evaluation System

The effects of the current teacher evaluation system are analyzed as follows based on existing empirical research findings, as well as other materials collected from the survey among 1360 teachers from 21 Beijing-based schools and follow-up supplementary interviews conducted by the author in the context of academic research (Li et al., 2018).

13.5.1 Does the Teacher Evaluation System Work?

Improving the overall teachers' credentials. Since the late 1970s, China has made significant progress in improving the credentials of its teachers, especially in terms of readjusting teachers' age structure and consolidating their educational background. By the end of 2000, the proportion of Chinese junior high school teachers holding an associate or college degree reached 72.9% and 14.25%, respectively, while such ratios for senior high school teachers reached 30.23% and 68.4%, indicating a dramatic growth compared with 1979 (Department of Development and Planning, Ministry of Education, 2001, pp. 64–65). The teacher evaluation system played a remarkable role in ensuring that in-service teachers could meet basic requirements. On the one hand, given that the professional standards for teaching credentials were not issued until 2012, for a long while, the evaluation of teacher qualifications served as criterion to measure whether a teacher was qualified for the teaching position. As a result, the incumbent teachers or those intended to become teachers were promoted to meet such basic requirements as attaining certain educational level and mastering relevant teaching competences. On the other hand, the periodic teacher appraisal provided key benchmarks for teacher training at that time, as those whose performance was rated "disqualified" had to attend competency-based teaching training or accept continued education to meet the lowest threshold for a teaching post. The appraisal forced teachers to upgrade their academic degrees to meet the requirements and urged governments to provide more support.

Providing stronger motivations for teachers. It is expected that teacher evaluation, especially periodic appraisal and performance evaluation, can raise teachers' enthusiasm in work and their motivation to improve teaching skills through material incentives. Against the backdrop of China's new round of curriculum reform, many teachers may be reluctant to change and move beyond their "comfort zone" as they fear to jump into the unknown (Wong, 2012). Adopting new textbooks and teaching methods requires teachers to devote more effort. But they are unlikely to support curriculum reform when their efforts do not pay off in a short term. According to empirical studies, performance evaluation can arouse teachers' passion for reform. A survey on 1906 teachers from more than 60 schools reveals that 61.1% respondents agreed that the performance evaluation for the teachers in compulsory education arouses their work enthusiasm (Fan & Fu, 2011). However, according to a survey conducted among 547 teachers in 43 schools, when asked about the teachers' attitude toward the pay-for-performance incentive program, only 12.3% of the respondents chose "satisfied" (Yang & Du, 2014). However, within a school, many problems still exist in the distribution of the merit pay for teachers within schools. Among them, the most noteworthy problems are as follows: (1) The differences in teachers' merit pay are controlled within a small range, making teachers less motivated; (2) the merit pay of school administrators is significantly higher than that of teachers, and the merit pay of those teaching Chinese, Math, and other "core subjects" is higher than those teaching the subjects that were given less weight in exams such as Sports, Music, and Painting (Lyu & He, 2011). Although this gap might be small, teachers' performance largely depends on whether they are undertaking administrative work, or teaching core subjects, which is unfair and might dampen the enthusiasm of some teachers.

Theoretically, when it comes to teachers' professional title evaluation and awards selection, in addition to material incentives, professional recognition and honors are also expected to motivate teachers to improve their teaching effectiveness. However, in reality those teachers with longer years of service always enjoy priorities in competing for professional titles and awards, which also discourages those relatively young teachers.

Optimizing school teacher personnel management. Some schools have created teacher evaluation system and developed a set of standards, procedures, and tools. According to a survey conducted by the author, 77.8% of the surveyed teachers agreed that "the teacher evaluation system is improving gradually". Regarding the fairness, openness, and impartiality of the assessment process, 71.1% of teachers held a positive attitude, and only 8.3% of teachers expressed a negative attitude (Li, 2016, p. 112). A variety of evaluation results have since been widely used in teacher's personnel management, especially in teachers' promotions, salaries, and training. Formerly, school leaders' opinions can decide a teacher's promotion, but now, evaluations on teachers' performance, honorary titles, and classroom teaching are also taken into consideration. However, in some schools, the use of evaluation results still has a lot of room to improve. As with the words of the two interviewed principals:

"The most important reason for designing such a procedure is to make everyone agree with the final decisions. This procedure is a must"; "I just want to use this data to push ahead of the initiative (Li, 2016, p. 113).". Hence, school administrators organize teacher evaluations not purely for improving management, but also with the intention to reduce teachers' doubts over the management as much as possible through a set of so-called standardized procedures and in a bid to avoid criticism due to the lack of evaluation system. In addition, due to the lack of professional support, many have raised doubts on whether the current teacher evaluation system can provide a solid basis for school management. On the one hand, schools prefer to conduct evaluations through reviewing written documents and rating by school authorities and teachers, which is not sufficient to reveal the situation of teachers in depth. On the other hand, even if some schools adopt a quantitative evaluation method, the evaluation tools may lack explicit indicators and necessary analysis such as reliability and validity.

Promoting teachers' professional development. The existing teacher evaluation system promotes teachers' professional development in three ways. First, some evaluations cover a great number of activities that promote teachers' professional development. For example, classroom observation is needed in the classroom teaching evaluation, of which the results may be used for high-stakes evaluation such as performance evaluation and professional tittle evaluation. Thus, teachers are motivated to conduct periodic classroom observation, which promotes teachers' in-depth cooperation and mutual learning. In addition, teacher evaluation prompts teachers to participate in the activities conducive to professional development. For example, teachers are required to participate in certain mandatory professional activities if they want to get the eligibility for qualification evaluation; many schools regulate clear requirements in their performance evaluations framework regarding the number and level of the professional development activities that teachers are supposed to participate in, the number of collective lesson preparations, the number and records of teaching summaries, and the results gained in pedagogical research. A teacher has pointed out in a study that

> Appraisal can press us so that we have to get done the things [as required in the appraisal system], like lesson planning and teaching reflections.... As they are checked regularly by the school [administrators], we have to deal with such things seriously. Now I feel I have made many improvements by doing them. ... If they were not checked, we would have taken a perfunctory attitude to these things. As such, we still remained in the status quo for a significant period. (Zhang & Ng, 2011, p. 576)

Finally, teacher evaluation can also provide reference for teachers' professional development. One of the most typical examples is classroom teaching evaluation, from which teachers can receive feedbacks of their teaching strengths and weaknesses from their peers and school administrators, thus improving their teaching strategies through a mutual learning approach.

However, some problems still exist in the current evaluation mechanism. For example, some teachers do a perfunctory job in the evaluation, as manifested by failing to give careful assessment during the classroom observations or simply copying the lecturers' teaching materials. In the interviews conducted by the author, some teachers said frankly: “I am so busy that I can’t spare too much time for classroom observations. Sometimes it is purely to complete the work.” “Teachers seldom give negative assessment. We are colleagues. How can we unleash sharp criticism against their teaching? Everyone understands that it is just a formality.” “The school stipulates the minimum amount of classroom observations a teacher needs to attend each semester. Everyone gets involved just out of a desire to complete the task. We rarely give a careful assessment, unless we are asked to do so (Li, 2016, p. 92).”

13.5.2 What Restricts the Effects of the Teacher Evaluation System?

Some problems concerning the evaluation content and indicators, evaluation methods and tools, and the use of evaluation results have created barriers to making the best use of teacher evaluation. The widely used evaluation systems have even deviated from the original dual orientation of serving schools’ management and teachers’ professional development. Some researchers pointed out the essential cause: Teacher evaluation was not achieving its fundamental goal of “serving educating people” (Xin, 2020; Zhong, 2020).

In terms of evaluation content and indicators, the evaluation of educating individuals was simplified to some quantitative outcome-oriented indicators, which fails to reflect the core purpose of education. First, school administrators often rank teachers based on single year’s students’ test scores, which ignores the holistic development of students. According to the survey conducted in Beijing, 56.1% of the respondents agreed that “test-based teacher evaluation methods are used in our school”, especially in urban schools at the middle school level (Li, 2016, p. 114). This outcome-oriented evaluation concept forces teachers to place nearly all of their emphasis on students’ test scores and their achievement on standardized tests organized by districts and the state, rather than promoting holistic development of students as the fundamental purpose of education. Second, the current teacher evaluation employs indicators that are easy to measure and ignores the complex process of teaching and learning. According to the performance evaluation plans collected from schools, serving as a class teacher, group leader or middle-level school administrator as well as workload (specific hours) combined influence teachers’ performance-based pay. Other evaluation indicators are dominated by such frequency statistics as the training sessions a teacher has attended. The indicators concerning how teachers set teaching goals, organize teaching content, choose teaching methods, and carry out evaluation are not within the scope of evaluation in many schools. Third, the current teacher evaluation system merely focuses on whether teachers can fulfill their responsibilities in their daily routine work (e.g., submitting materials in a timely manner, carefully checking, and correcting homework), while teachers’ exploration of innovative methods in classroom teaching is neglected to a large extent. In the words of an interviewed teacher:

> I want to use new teaching methods, but I cannot guarantee that my students can improve their exam performance within a short period, and even their scores risk declining. But other teachers who use cramming methods can help their students improve test scores. The final teacher evaluation will not take the possible consequence of using my new teaching methods into account. So, who would like to change? (Li, 2016, p. 81)

In terms of evaluation methods, there is still a scarcity of diversified evaluation methods in implementation, thus failing to demonstrate teachers' efforts in educating people. First, the current evaluation mechanism places too much emphasis on quantitative statistics and neglects the use of qualitative methods. Admittedly, quantitative statistics on teachers' work can provide school administrators with clear standards to follow when measuring whether teachers meet the prescribed requirements, boosting its transparency and fairness. However, it cannot depict the whole picture of teachers' education process as qualitative methods do, let alone reflecting teachers' educational philosophies and the changes they have made. Second, school administrators mostly focus on students' final academic achievements, but overlook their original academic foundation, family background, and other preexisting factors. Such an evaluation directs teachers to favor students with better academic performance. Third, although some new evaluation methods have been employed in the evaluation system, they have not yet received sufficient attention. For example, classroom observation becomes increasingly popular in recent years, but it is often only regarded as a method of teacher research and training and has not truly become an important yardstick to measure teacher performance in the evaluation. Teachers consider that classroom observation is not as objective as test scores, thus should not be a determinant of their salaries.

Regarding the use of evaluation results, in spite of the many types of teacher evaluations, limited effects have been seen in promoting teachers to improve their teaching. Apart from the fact that the current evaluation system fails to reflect the actual teaching practice, there are two other crucial issues. First, teachers' acceptance of the evaluation results is currently at a low level due to their inadequate involvement in the whole evaluation process. On the one hand, teachers, especially those without titles, have limited opportunities to participate in formulating evaluation content and indicators. In some teachers' opinions, the contents and indicators essential to improve education and teaching practices are not included in the evaluation, hence no resonance with the existing indicators and standards. On the other hand, ordinary teachers lack the opportunity to participate in the evaluation process, with self-evaluation playing negligible roles, and only teachers with manger roles or titles have the opportunity to participate in the peer evaluation. Lacking opportunities for self-clarification and self-reflection reduces teachers' recognition of the evaluation process as a fair one. The survey conducted among the 841 elementary and middle school teachers in Fujian province revealed that 33.5% of the surveyed teachers thought that the current teacher evaluation system took teacher self-evaluation into account and 35.9% thought the system did not consider teacher self-evaluation (Liang, 2012). Second, a lack of timely feedback on evaluation results and the guidance for improving teaching competence impacts as well. The survey shows that some schools do not provide feedback on the results of the evaluation at all, including the specific results of performance appraisal and the ratings given by students. Consequently, guiding teachers to improve teaching practice based on the evaluation results is unlikely to happen.

Finally, repeated evaluations have taken up teachers' enormous time and energy, burning them out with increasing non-teaching workloads. Overwhelmingly, teachers need to go through various evaluations such as classroom teaching evaluation, performance evaluation, and professional title evaluation. The requirements for the evaluation process and material preparation are fairly strict and tedious. Taking the professional title evaluation in Guangdong (a province in southeastern China) for example, apart from filling out forms and participating in interviews, applicants for the first-grade title still need to submit 13 types of documents in print and electronic editions (see Table 13.4). At present, most teachers have been besieged with such problems as long working hours, heavy teaching tasks, and a plethora of issues unrelated to teaching. In an online survey involving more than 100,000 elementary and middle school teachers nationwide, nearly 70% of the teachers agreed that "teaching is an exhausting work" (Xiong & Jiang, 2019). At present, a large number of evaluation activities have not been effectively integrated, which can easily impose extra burdens on teachers, especially increasing teachers' time and energy input in writing and preparation. This online survey also showed that the teacher evaluation mechanism mentioned above was the second most stressful source for teachers' pressure, followed by the government officials' frequent inspections in schools and the corresponding evaluations.

13.6 Future Directions of China's Teacher Evaluation System

The *Overall Plan for Deepening Educational Evaluation Reform in the New Era*, a decision on education evaluation reform made by China's top policymakers, calls for improving the teacher evaluation to facilitate the nurturing of young generations (Central Committee of the CPC & State Council, 2020). In general, the reform of teacher evaluation in the new era is expected to overcome the phenomenon of overemphasizing imparting knowledge rather than cultivating people and to establish a system that guides teachers to be committed to teaching and education for younger generations. Specifically, the initiative can be fueled from the following perspectives.

First, the reform of teacher evaluation should coordinate the dual functions of teacher evaluation in teacher personnel management and promoting teachers' professional development and adhere to the ultimate goal of educating people. In terms of the management function of the evaluation, the focus of the reform is to use more diverse and appropriate evaluation methods to evaluate teachers' work and ensure that the final evaluation results are more convincing and objective, in a bid to motivate teachers to devote themselves to the fundamental task of educating people. As for the function of promoting the professional development of teachers, the focus of the

Table 13.4 Documents submitted when applying for first-grade title in Guangdong

No.	Documents submitted
1	Scanning copy of personal photo and ID card
2	Academic and degree certificates
3	Social security credentials
4	Results of national professional and technical personnel of foreign language grade title examination (or certificate of exempt from examination), and the national professional and technical personnel computer application ability test (or certificate of exempt from examination)
5	Certificate of completion of continuing education and training
6	Annual performance evaluation rating scale
7	Honors (such as "excellent head teacher"), summary of education experience and case study
8	Demonstration of teaching open classes, personal teaching features, and mentor of prizewinners, teaching reflections and notes, honorary certificates
9	Papers, works, works of translation
10	Teacher qualification certificate, professional expertise qualification certificate, letter of appointment
11	Papers presented on academic conferences
12	Demonstration of participation in research projects
13	Honor certificates, reports on work, teaching and research achievements, papers

Note Adapted from "Notice on Elementary and Secondary School Teachers' Professional Title Evaluation", by Department of Human Resources and Social Security of Guangdong Province and Department of Education of Guangdong Province (2018)

reform is to clarify the requirements for teachers in educating people through evaluation, make the evaluation results more scientific and acceptable, so that teachers can improve their teaching based on the evaluation results.

Second, more focus should be directed to the process and actual effectiveness of educating people as the core elements of teacher evaluation. The indicators in various teacher evaluation mechanisms need to be restructured around the theme of educating people. (1) It is necessary to correct the tendency of using test-oriented teacher evaluation methods and place less emphasis on promotion rates. On the one hand, the government at all levels shall not use promotion rates to evaluate the performance of schools and teachers; on the other hand, the phenomenon that the overdependence of outcome-oriented evaluation (e.g., based on student test scores, including those in certain critical examinations) should be avoided, and the weight of such test-based indicators should be lowered down. Undoubtedly, such action needs to abide by policies and follow teachers' advice. (2) It is important to strengthen the evaluation of the process of educating students, teacher ethics, teaching style, as well as whether teachers can promote the moral, intellectual, physical, and artistic development of students, curriculum development, and provision of academic guidance to students.

(3) The innovation of educating students should be incorporated into the teacher evaluation framework. The weight of teaching improvements and innovations should be increased in teacher evaluation. It is noteworthy that evaluators should concern about improvements and innovations in daily classes rather than in teaching competitions.

Third, the evaluation methods should be optimized to increase evaluation accuracy. The improvement of the outcome-oriented evaluation should be based upon both quantitative and qualitative methods, including statistical methods, review of written teaching records, student–teacher discussions, students' ratings, and classroom observations. A more comprehensive evaluation system is supposed to be devised to assess teaching effectiveness rather than heavily depending on quantitative statistics. Then process-oriented evaluation should be enhanced with classroom observation, portfolio assessment, and other comprehensive methods to evaluate the process of educating students. With this approach, evaluation can better demonstrate the complexity of the teaching process and provide more information for teachers to improve their teaching. Governments can encourage researchers or third-party evaluation agencies to conduct this type of research and training through commissioned research or service procurement and provide tools and online platforms. Furthermore, value-added evaluation should be explored to evaluate teachers. Schools should not evaluate teachers by simply looking at scores of final examinations, *Zhongkao* or *Gaokao*, rather, they should give more weight to teachers' contribution to students' progress. Although value-added evaluation has led the way for Chinese education evaluation reform, extensive researches need to be carried out concerning the use of value-added evaluation models and interpretation of results. Current studies have not reached a consensus on the stability of such value-added models and whether it is really effective to distinguish effective and less-effective teachers (Darling-Hammond et al., 2012).

Fourth, a greater emphasis should be placed on the use of the evaluation results to help teachers fulfill the fundamental task of educating people. More teachers should be invited to participate in the evaluation work in the first place. They are supposed to be encouraged to conduct self-reflection and evaluation and provide peer assistance and expert guidance for their self-evaluation. Furthermore, teachers should have channels to voice their opinions through teacher plenary meetings, teacher representative meetings, work teams, surveys, and so forth. Then the feedback guidance on evaluation results should be enhanced. In terms of the evaluation results to be used in teacher management, such as performance appraisal and professional title evaluation, schools not only need to publicize evaluation results within the school, but also disclose the information related to the shortcomings of the teaching practice to individual targeted teachers to help them enhance their teaching competence. In terms of the evaluation results that will have a close bearing on teachers' professional development, such as classroom teaching evaluation, schools must form their classroom observation frameworks and procedures based on existing research findings and other schools' experience, thus laying a solid foundation for collecting more evidence. Both teachers and school administrators should take an active part in the class observations and offer tailored guidance to help teachers improve teaching practices.

Fifth, governments at all levels need to merge various evaluations. For one thing, governments should simplify the evaluation process, especially reducing submitted documents. For example, an applicant only needs to submit limited representative documents, which exceeds the minimum number of characters required. For another, more emphasis should be put on integrating and sharing various information and evaluation results. For instance, evaluation results for classroom teaching can be regarded as evidence for performance evaluation, whose results can serve as the basis for evaluation of professional titles, honors, and awards. Meanwhile, governments can create platforms to collect evaluation information of various kinds of teacher evaluation, thus reducing repetition in the collection.

Sixth, researchers need to strengthen research on China's teacher evaluation system. Despite the fact that an integrated teacher evaluation system has been established, there is still a lack of profound empirical studies on the system's effectiveness. Domestic studies tend to focus on performance evaluations, such as governments' financial support to performance pay, development and implementation of schools' performance pay scheme, and impact of performance pay evaluation on teachers (e.g., Cai et al., 2018; Jiang et al., 2014). Also, some researches investigate classroom teaching evaluations, but the emphasis is on its role and implementation, rather than effectiveness (e.g., Cui, 2012; Cui et al., 2013; Shen & Cui, 2008). Researches on the evaluation of teachers' qualifications, professional titles, awards, and honors are scarce, not to mention empirical studies on their effectiveness.

As such, three suggestions are provided for the studies in the future: (1) Researches should focus on the effectiveness of evaluation of teachers' qualifications, professional titles, awards, and honors. Key factors that affect such effectiveness should be identified to provide a basis for the improvement of evaluations. (2) China needs to improve its research methodology in teacher evaluation. Since curriculum reform, there have been a growing number of theoretical analyses and reviews of teacher evaluation in China, but these are not empirical studies that provide solid evidence for researchers and practitioners. More empirical studies are in need to investigate the topics more thoroughly, such as examining the relationship between orientation, content, method of teacher evaluation and teachers' satisfaction, organizational identification, self-efficacy, and students' achievements. (3) More studies on the content and methods of evaluation should be conducted. Under the current evaluation systems, this type of studies is scarce, which makes it hard for schools to adopt real scientific indicators based upon solid studies to evaluate teachers' performance. For example, issues that deserve researchers' attention and solutions include aspects regarding classroom teaching observation, and moreover, the employment of qualitative evaluation methods in high-stakes evaluations.

Note

1. Reform and opening up refers to the historical period in which China implemented a series of policies of domestic reform and opening up since the 3rd Plenary Session of the 11th Central

Committee of the Chinese Communist Party in December 1978. Domestically, China gradually established the socialist market economy by contracting rural collective land to farmers and increasing the autonomy of enterprises. Opening up to the outside world was achieved through the establishment of special economic zones to promote international exchanges and trade. Since the reform and opening up, China has witnessed rapid social and economic development, with the level of industrialization, urbanization, and internationalization rising significantly. Therefore, people tend to regard it as an important backdrop to study the reform and development of various undertakings in China.

References

Cai, Y., Bi, Y., Wang, L., Cravens, X. C., & Li, Y. (2018). The construct of teachers' pay satisfaction: A case study of primary and secondary schools in China. *Teachers and Teaching: Theory and Practice*, *24*(4), 431–449. https://doi.org/10.1080/13540602.2017.1421163

Carney, S. (2009). Negotiating policy in an age of globalization: Exploring educational "policyscapes" in Denmark, Nepal, and China. *Comparative Education Review, 53*(1), 63–88. https://doi.org/10.1086/593152

Central Committee of the Communist Party of China. (1985, May 27). *Seminal decision on reform of educational system.* http://www.moe.gov.cn/jyb_sjzl/moe_177/tnull_2482.html (in Chinese).

Central Committee of the Communist Party of the People's Republic of China, State Council of the People's Republic of China. (2020, October 13). *Overall plan for deepening educational evaluation reform in the new era.* http://www.gov.cn/zhengce/2020-10/13/content_5551032.htm (in Chinese).

Chu, H. (2008). A brief comment on the 30-year reform of China's basic education administration system. *School Administration, 11*, 4–8. (in Chinese).

Cui, Y. (2012). On the paradigm of LICC: A new way of professional classroom observation. *Educational Research, 33*(5), 79–83.

Cui, Y., Shen, Y., & Wu, J. (2013). *Classroom observation II: Towards professional "Tingpingke."* East China Normal University Press. (in Chinese).

Darling-Hammond, L. (2012). The right start: Creating a strong foundation for the teaching career. *Phi Delta Kappan, 94*(3), 8–13. https://doi.org/10.1177/003172171209400303

Darling-Hammond, L., Amerin-Beardsley, A., Haertel, E., & Rothstein, J. (2012). Evaluating teacher evaluation. Phi Delta *Kappan*, *93*(6), 8–15.

Department of Human Resources and Social Security of Beijing. (2016, April 18). *Requirements of applying for teacher professional title in elementary and secondary schools in Beijing*. http://www.beijing.gov.cn/zhengce/zhengcefagui/201905/t20190522_59164.html (in Chinese).

Department of Human Resources and Social Security of Guangdong Province, & Department of Education of Guangdong Province. (2018, October 25). *Notice on elementary and secondary school teachers' professional title evaluation.* https://www.gdhrss.gov.cn/gsgg/14382.jhtml (in Chinese).

Department of Planning, Ministry of Education. (1980). *China education statistical yearbook (1979)* (Unpublished internal documents). Beijing (in Chinese).

Department of Planning, Ministry of Education. (2001). *China education statistical yearbook (2000).* People's Education Press (in Chinese).

Department of Teacher Education, Ministry of Education of the People's Republic of China, & National Education Examinations Authority of the People's Republic of China. (2011, October 11). S*tandards for the teacher qualification examination.* http://ntce.neea.edu.cn/html1/category/1511/692-1.htm (in Chinese).

Department of Teacher Education, Ministry of Education of the People's Republic of China, & National Education Examinations Authority of the People's Republic of China. (2012a, May

9). *Outlines for tests of the teacher qualification examination.* http://ntce.neea.edu.cn/html1/category/1507/1099-1.htm (in Chinese).

Department of Teacher Education, Ministry of Education of the People's Republic of China, & National Education Examinations Authority of the People's Republic of China. (2012b, May 9). *Outlines for trial lecture of the teacher qualification examination.* http://ntce.neea.edu.cn/html1/category/1511/693-1.htm (in Chinese).

Evans, L. (2013). The 'shape' of teacher professionalism in England: Professional standards, performance management, professional development and the changes proposed in the 2010 White Paper. *British Educational Research Journal, 37*(5), 851–870. https://doi.org/10.1080/01411926.2011.607231

Fan, X., & Fu, W. (2011). Teacher performance salary reform in compulsory education: Background, effectiveness, problems and countermeasures: Based on a survey of 32 counties (cities) in 4 provinces in central China. *Journal of Huazhong Normal University (Humanities and Social Sciences), 50*(6), 128–137. https://doi.org/10.3969/j.issn.1000-2456.2011.06.019 (in Chinese).

Huang, F. (2004). Curriculum reform in contemporary China: Seven goals and six strategies. *Journal of Curriculum Studies, 36*(1), 101–115. https://doi.org/10.1080/002202703200004742000174126

Jiang, J., & Du, Y. (2014). On the project design of enhancing financial input of primary and secondary school teachers' salary and its feasibility analysis. *Educational Research, 419*(12), 54–60. (in Chinese).

Lai, M. H., & Lo, L. N. K. (2007). Teacher professionalism in educational reform: The experiences of Hong Kong and Shanghai. *Compare, 37*(1), 53–68. https://doi.org/10.1080/030579206010

Li, G. (2016). *A study on the changes of management of teachers' instructional improvement* (Unpublished doctoral dissertation). Beijing Normal University, China (in Chinese).

Li, T., Jin, C., & Jin, Z. (2018). How is the effect of the reform of primary and secondary school teachers' professional title system: A policy evaluation study based on multiple evaluation theory. *Research in Educational Development*, *38*(18), 17–23. https://doi.org/10.14121/j.cnki.1008-3855.2018.18.005(in Chinese).

Liang, F. (2012). Survey report of teacher performance evaluation in Fujian Province. In Y. Wang & S. Lin (Eds.), *Research of teacher performance evaluation* (pp. 40–54). Shanghai People's Publishing House. (in Chinese).

Liu, S., & Teddlie, C. (2003). The ongoing development of teacher evaluation and curriculum reform in the People's Republic of China. *Journal of Personnel Evaluation in Education, 17*(3), 243–261. https://doi.org/10.1007/s11092-005-2982-x

Liu, S., & Zhao, D. (2013). Teacher evaluation in China: Latest trends and future directions. *Educational Assessment Evaluation & Accountability, 25*(3), 231–250. https://doi.org/10.1007/s11092-013-9168-8

Liu, S., Xu, X., & Stronge, J. H. (2016). Chinese middle school teachers' preferences regarding performance evaluation measures. *Educational Assessment Evaluation & Accountability, 28*(2), 161–177. https://doi.org/10.1007/s11092-016-9237-x

Lv, Yu Y., & He, Z. (2011). Surveys of teacher performance evaluation in Shanxi Province. *Education Exploration*, *237*(3), 73–74 (in Chinese).

Ministry of Education of the People's Republic of China, Central Organization Department of the CPC, Central Propaganda Department of the CPC, Central Office of the CPC, National Development and Reform Commission of the People's Republic of China, Ministry of Public Security of the People's Republic of China, ... Ministry of Human Resources and Social Security People's Republic of China. of the (2020, September 22). *Several opinions of the Ministry of Education and other seven departments on further stimulating the vitality of elementary and secondary schools.* http://www.moe.gov.cn/srcsite/A06/s3321/202009/t20200923_490107.html (in Chinese).

Ministry of Education of the People's Republic of China. (2001, June 8). *Guideline on the reform of curriculum in basic education.* http://www.moe.gov.cn/jyb_sjzl/moe_364/moe_302/moe_309/tnull_4672.html (in Chinese).

Ministry of Education of the People's Republic of China. (2002, December 27). *Notification on advancing reform of evaluation and exam system for elementary and secondary schools*. http://www.moe.gov.cn/srcsite/A26/s7054/200212/t20021218_78509.html (in Chinese).

Ministry of Education of the People's Republic of China. (2003, October 21). *Notice on further strengthening the management and professional morality education of elementary and secondary school teachers*. http://www.moe.gov.cn/jyb_xxgk/gk_gbgg/moe_0/moe_9/moe_40/tnull_144.html (in Chinese).

Ministry of Education of the People's Republic of China. (2008, December 31). *Guiding opinions of the ministry of education on conducting teacher performance evaluation in compulsory education schools*. http://old.moe.gov.cn/publicfiles/business/htmlfiles/moe/s7051/201412/xxgk_180682.html (in Chinese).

Ministry of Education of the People's Republic of China. (2013, August 15). *Interim measures for regular registration of teacher qualifications in elementary and secondary schools*. http://ntce.neea.edu.cn/html1/report/1507/1140-1.htm (in Chinese).

Ministry of Education the People's Republic of China. (2020a, May 20). *Statistical bulletin on national education development in 2019*. http://www.moe.gov.cn/jyb_sjzl/sjzl_fztjgb/202005/t20200520_456751.html (in Chinese).

Ministry of Education of the People's Republic of China. (2020b, June 11). *Basic situation of national education in 2019*. http://www.moe.gov.cn/jyb_sjzl/moe_560/jytjsj_2019/ (in Chinese).

Ministry of Human Resources and Social Security, & Ministry of Education of the People's Republic of China. (2015, August 18). *Basic standards and requirements for teachers' professional title evaluation in elementary and secondary schools*. http://www.moe.gov.cn/jyb_xxgk/moe_1777/moe_1779/201509/t20150902_205165.html (in Chinese).

National People's Congress of the People's Republic of China. (1986, April 12). *Compulsory education law*. http://www.moe.gov.cn/s78/A02/zfs_left/s5911/moe_619/201001/t20100129_15687.html (in Chinese).

Organization for Economic Co-operation and Development. (2013). *Synergies for better learning: An international perspective on evaluation and assessment*. OECD Publishing. https://doi.org/10.1787/9789264190658-en

Organization for Economic Co-operation and Development. (2020). *Benchmarking the performance of China's education system*. OECD Publishing. https://doi.org/10.1787/4ab33702-en

Shen, Y., & Cui, Y. (2008). *Classroom observation: Towards professional "Tingpingke."* East China Normal University Press. (in Chinese).

Standing Committee of the National People's Congress of the People's Republic of China. (1993, October 31). *Teacher act*. http://www.moe.gov.cn/s78/A02/zfs_left/s5911/moe_619/tnull_1314.html (in Chinese).

Standing Committee of the National People's Congress of the People's Republic of China. (2018, December 29). *Non-state education promotion law*. http://www.npc.gov.cn/npc/c30834/201901/8c8f598f14ba4728a6181aec8cb1b90a.shtml (in Chinese).

State Council of the People's Republic of China. (1995, December 12). *Regulations on the qualifications of teachers*. http://www.moe.gov.cn/s78/A02/zfs_left/s5911/moe_620/tnull_3178.html (in Chinese).

State Council of the People's Republic of China. (2008, December 21). *Guiding opinions on implementing performance-based pay in compulsory education schools*. http://www.gov.cn/ldhd/2008-12/21/content_1184109.htm (in Chinese).

State Education Commission of the People's Republic of China. (1983, August 22). *Recommendations on strengthening administration of teachers in elementary and secondary schools*. https://www.pkulaw.com/CLI.4.65901 (in Chinese).

State Education Commission of the People's Republic of China. (1986a, September 6). *Interim measures for qualification certificates for elementary and secondary school teachers*. https://www.pkulaw.com/CLI.4.41358 (in Chinese).

State Education Commission of the People's Republic of China. (1986b, May 19). *Interim rules for the duties of elementary school teachers and interim rules for the duties of middle school teachers*. https://www.pkulaw.com/CLI.5.38373 (in Chinese).

Stronge, J. H. (2006). Teacher evaluation and school improvement: Improving the educational landscape. In J. H. Stronge (Ed.), *Evaluating teaching: A guide to current thinking and best practice* (2nd ed., pp. 1–24). Corwin Press.

Taut, S., Santelices, M. V., Araya, C., & Manzi, J. (2011). Perceived effects and uses of the national teacher evaluation system in Chilean elementary schools. *Studies in Educational Evaluation, 37*(4), 218–229. https://doi.org/10.1016/j.stueduc.2011.08.002

Wang, J. (2021, March, 3). *The central committee of the democratic progressive party: The degree level of teachers employed in elementary and secondary schools in China is expected to be improved*. https://www.thepaper.cn/newsDetail_forward_11533791 (in Chinese).

Wise, A. E., Darling-Hammond, L., McLaughlin, M. W., & Bernstein, H. T. (1985). Teacher evaluation: A study of effective practices. *The Elementary School Journal, 86*(1), 60–121. https://doi.org/10.1086/461437

Wong, J. L. N. (2012). How has recent curriculum reform in China influenced school-based teacher learning? An ethnographic study of two subject departments in Shanghai, China. *Asia-Pacific Journal of Teacher Education*, *40*(4), 347–361.https://doi.org/10.1080/1359866X.2012.724654

Xin, T. (2020, November 11). Educating people: The fundamental functions of education evaluation. *China Education Daily*, p. 5.

Xiong, J., & Jiang, B. (2019). Investigation on the current situation of primary and middle school teachers' workload and corresponding countermeasures. *Teacher's Journal, 9*, 72–75. (in Chinese).

Yang, X., & Du, X. (2014). The Influence factors and attribution analysis about incentive effects of the teacher performance pay in compulsory education: Based on the Investigation in Sichuan Province. *Basic Education*, *11*(4), 32–41.https://doi.org/10.3969/j.issn.1005-2232.2014.04.005(in Chinese).

Zhang, B. (2011). Analysis of the deviation behaviors in carrying out the policies for evaluating professional title for primary and middle school teachers. *Educational Academic Monthly, (01)*, 81–83. https://doi.org/10.16477/j.cnki.issn1674-2311.2011.01.015(in Chinese).

Zhang, X. F., & Ng, H. M. (2011). A case study of teacher appraisal in Shanghai, China: In relation to teacher professional development. *Asia Pacific Education Review, 12*, 569–580. https://doi.org/10.1007/s12564-011-9159-8

Zhao, H., Hui, X., & Fu, C. (2011). A research on the present situation of performance-based pay for compulsory education teachers in China: Based on a study of 279 schools in 77 counties of 25 provinces. *Theory and Practice of Education, 31*(10), 24–27. (in Chinese).

Zhong, B (2020, October 19). Comprehensively build an evaluation system for educating people in the new era. *China Youth Daily*, p. 8.

Chapter 14
Teacher Evaluation System in South Korea

Jisung Yoo

Abstract This chapter presents an overview of the national standardized teacher evaluation system in South Korea, including its development, purpose, design, and implementation. Specifically, the antecedents of the current system are discussed, aiming to provide an understanding of the evolution of teacher evaluation in Korea and the challenges encountered. Also discussed is the influence of various political actors such as the government, teacher unions, media, and public perceptions, most notably the public's loss of trust in the education system due to the phenomenon known as the "classroom collapse." Components of the evaluation system are described, with a focus on Korea's unique assessment of teacher performance and evaluation consequences (i.e., sabbaticals as rewards and professional development for improvement). The chapter concludes with a discussion of the perceived effectiveness of the current evaluation system in terms of achieving its ultimate goals: improving student achievement and ensuring the equitable distribution of high-quality education throughout the nation.

14.1 Introduction

Teacher evaluation can be a tool to monitor and ensure the high performance of teachers, which in turn can serve as a means of ensuring the equitable distribution of high-quality education to all students. South Korea has enacted educational policies, developed under different government administrations over the past 50 years, to address the inequitable distribution of high-quality education, especially in rural and low-socioeconomic (SES) areas (Choi & Park, 2016; Seo, 2012). As scholars have emphasized the important role that teachers play in impacting student achievement, especially because they are major actors involved in the students' education (Goldhaber, 2002; Rivkin et al., 2005; Rockoff, 2004), a teacher evaluation system

J. Yoo (✉)
Konkuk University Glocal Campus, Chungju, South Korea
e-mail: jisyoo@kku.ac.kr

J. Manzi et al. (eds.), *Teacher Evaluation Around the World*,
Teacher Education, Learning Innovation and Accountability,
https://doi.org/10.1007/978-3-031-13639-9_14

provides a means of assessing teacher performance and identifying highly effective teachers while providing professional development opportunities for others who need improvement, which can ultimately result in improving student achievement.

Before examining Korea's current teacher evaluation system, it is important to understand some differences and perhaps unique aspects of the Korean culture and educational system. First, students with little financial support have faced difficulty entering the best universities, including Seoul National University, Yonsei University, and Korea University. Students who have had the advantage of taking expensive after-school lessons in private institutions—a popular and common practice among Korean families that can afford the tuition—have more possibilities of entering those prestigious schools. This disparity of financial resources results in inequalities in educational opportunities. According to the 2012 report released by the Ministry of Education, Science, and Technology (MEST), the university admission rate for students from high schools in high socioeconomic status (SES) areas is almost twice that of students from low-SES high schools (Seo, 2012).

Korea's educational system also differs from systems in most other countries in that public school teachers are permanently employed. They obtain this permanent status upon earning their teaching certificate, passing the national teacher examination, and securing employment in a public school (Organisation for Economic Co-operation and Development, 2009a, 2009b). Public school teachers are assigned to a different school every five years—a rotation system that aims to avoid corruption and other issues that can result from long-term employment at the same school. As Korean public school teachers have secure employment and do not need to be concerned about dismissal as a consequence of teacher evaluation, teacher evaluation has traditionally been used only for the purpose of promoting teachers to the few administrative positions (Kang, 2013).

Developing an effective teacher evaluation system in Korea was heavily influenced by the interest in addressing public demand for school accountability (Yoo, 2009). This interest and concern reached a critical point after a media report claimed that the public school education system was not functional. This phenomenon was described as "school collapse" or "classroom collapse," a term coined by the media to describe a social phenomenon in which teachers are unable to manage student behaviors, such as various kinds of students' disengagement, including sleeping, playing games, teasing peers, chatting, moving around, and ignoring teacher questions or directions in lessons (Whang et al., 2001). A heated public discourse on this issue, as well as teachers' strong criticism of the traditional teacher evaluation system, led to subsequent school reform efforts, including the development of the current teacher evaluation system by the South Korea government (Kang, 2013).

Korea's teacher evaluation system also differs from the systems used in some other countries, such as the USA, in that it is a national standardized system implemented in all 17 regions of the country: Seoul, Busan, Daegu, Incheon, Gwangju, Daejeon, Ulsan, North Chungcheong, South Chungcheong, Gangwon, Gyeonggi, North Gyeongsang, South Gyeongsang, North Jeolla, South Jeolla, Sejong, and Jeju

Island. However, the differences among the rural and urban regions, especially differences in teacher salaries and working conditions, have made it challenging to maintain a national teacher evaluation system that ensures equitable distribution of highly effective teachers in all schools throughout the nation.

The teacher evaluation system in Korea has evolved over the past six decades. As discussed in the next section, Korea has developed and implemented three national teacher evaluation systems, each focused on a different consequence of evaluation but all aimed to improve student achievement and ensure equitable distribution of high-quality teachers in all regions of the country.

14.2 Evolution of the Teacher Evaluation System in Korea

Korea has developed and implemented three teacher evaluation systems since 1964. The first two systems were based on promotion and merit pay, respectively. The third and current system is based on rewarding high-performing teachers and encouraging further professional development for those who need improvement. The directives of the three teacher evaluation policies are detailed below.

14.2.1 Teacher Evaluation System for Promotion

In 1964, a teacher evaluation system was adopted for the sole purpose of promotion based on teacher performance ratings. The aim of teacher performance ratings was to ensure fair and objective promotions (Choi & Park, 2016). Teacher performance ratings for promotion targeted two groups: teachers and vice-principals. However, several aspects, such as the evaluation areas and evaluators, were different. This evaluation policy required teachers and vice-principals to submit an annual self-report on their performance. These self-reports were assessed by multiple evaluators, including the principal, the vice-principal, and three or more peer teachers, with weightings of 30%, 40%, and 30%, respectively. The norm-referenced evaluation, based on a total possible score of 100 points, was designed to compare and rank teachers in relation to one another (Choi & Park, 2016).

Two areas of performance were the focus of the teacher performance ratings for this promotion system: qualification and attitude, and work performance. Attitude was assessed based on characteristics as an educator and attitude as a public official, while the assessment of teacher performance focused on instruction, student guidance, and educational research and administrative service (Choi & Park, 2016).

A report by the Organisation for Economic Co-operation and Development (OECD) criticized that Korea's first teacher evaluation system, based on promotion, had several critical problems (Coolahan et al., 2004). One of the most serious problems was the promotion system failed to provide incentives for professional development throughout a teacher's career, and no systematic arrangement of rewards was included to recognize excellence in teaching. Another issue was the lack of validity in that the promotion system emphasized years of teaching, thereby excluding young and able teachers (Jeon, 2001). Therefore, this evaluation system focused only on experienced teachers, with evaluation of academic instruction comprising only 16% of the performance scores. Furthermore, the evaluation results were not even open for review (Kang, 2013). Thus, this evaluation system was ineffective due to the lack of constructive feedback, suggestions, or professional development opportunities and other essential assessments to ensure teacher accountability in classroom teaching and provide rewards for excellence in teaching.

14.2.2 Teacher Evaluation System for Performance-Based Pay

The financial crisis in Korea in the late 1990s led to the Korean government developing a new evaluation system with the purpose of promoting a creative and performance-based work environment for public officials, including teachers (Choi & Park, 2016). The performance-based pay system for public education teachers was introduced in 2001. General guidelines for teacher evaluation were issued by Korea's Ministry of Education. Although standards for performance evaluation varied across schools, they included the evaluation areas of instruction, student guidance, administrative service, and professional development (Choi & Park, 2016).

This system was implemented to encourage constructive competition among teachers and reward high-quality teachers with merit pay (Ministry of Education, 2012). As this evaluation system was designed to reward high-performing teachers and provide an incentive for other teachers to improve, 90% of a teacher's remuneration was based on performance, with the remaining 10% being evenly distributed. However, in response to teacher protests in 2002, this ratio was adjusted to 10% of a teacher's remuneration based on performance, with 90% evenly distributed. This pay ratio was later adjusted again to increase remuneration based on performance to 50% (Choi & Park, 2016; Seo, 2012).

As the monetary reward was not significant, it was an ineffective incentive to teachers. Furthermore, this system was ineffective because teachers knew their jobs were secure and they were generally already content with their high salaries (Kang, 2001). Consequently, this school performance-based pay system was abolished, and reform efforts were made to develop a new teacher evaluation system (Ministry of Education, 2012). It should be noted that although pay for performance failed to motivate teachers to improve their teaching in the context of the Korean educational system, other possible benefits of pay for performance may have been overlooked. For example, a recent US study by Pham et al. (2021), based on a meta-analysis of the findings of 44 primary studies, found that having a merit pay program is associated with a modest, statistically significant, positive effect on student test scores.

In 2010, the term "classroom collapse" was introduced in the discourse on education in South Korea to refer to the inability of teachers to teach due to disruptive student behaviors. Many of these behaviors were caused by the fact that many students were bored with classroom instruction in their public school, since they had already received similar instruction in private institutions. The economic structural inequality in South Korea allowed wealthy parents to send their children to private after-school institutes, while preventing less-wealthy parents to give their children the same advantage. It should be noted that parents consider that the quality of education provided by such institutes is much higher than that provided by public schools. Therefore, both the economic inequality and poor quality of the public school system were partly responsible for the "classroom collapse." This failure to control the classroom resulted in a situation where classroom lessons could not be delivered, the teacher's authority was threatened, and the basic function of schooling was weakened (Kang, 2013). This situation caused the public to lose trust in the education system.

In addition, Park (2006) described the inappropriateness of the current educational system, arguing that the traditional exam-based education system was inflexible and irrelevant to students' lives and simply focused on knowledge transmission rather than knowledge that is inseparable from students and their daily lives in a lifelong endeavor (Polanyi, 1958). Schools were criticized for not adapting to changes in the teenage culture, and, therefore, this school experience did not engage students (Kang, 2013). As classroom collapse represented the failure of public schools, it was clear that extensive change in all aspects of schooling was needed. A heated public discourse on the issue in Korea gained the attention of political actors, including Congress and the administration, which led to subsequent school reform efforts, including the development of the new, and current, teacher evaluation system (Kang, 2013).

14.2.3 Teacher Evaluation for Professional Development

In 2004, the Ministry of Education announced that a new teacher evaluation system would be developed with the main purpose of the professional development of teachers and the reduction of private tutoring expenditures. This teacher evaluation system, the third system proposed by the Ministry of Education, was implemented in selected schools in 2005 with specific directives for professional development and was fully implemented in all Korean schools in 2011. This system, which is the current national system, aims to develop teachers' skills and abilities, give productive feedback to teachers, and provide training programs (Ministry of Education & Human Resource Development, 2006).

In addition, in order to elicit 360-degree feedback, students and their parents, as well as principals, vice-principals and peer teachers, participate in the process as evaluators (Choi & Park, 2016; Seo, 2012). Furthermore, three groups of stakeholders evaluate all teachers in order to ensure the concreteness of the results. The first group

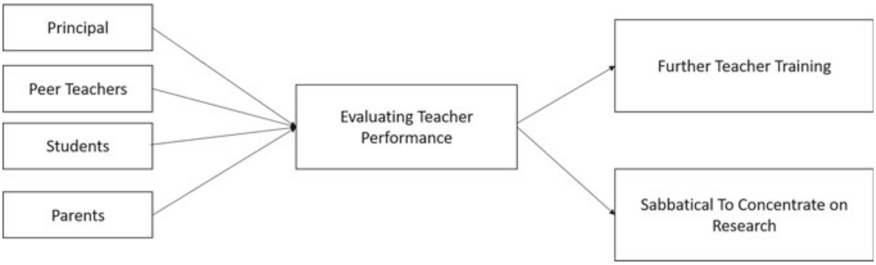

Principal
Peer Teachers
Students
Parents
Evaluating Teacher Performance
Further Teacher Training
Sabbatical To Concentrate on Research

Fig. 14.1 Conceptual model of the new teacher evaluation system in South Korea

comprises more than five peer teachers, including at least either the principal or vice-principal and at least either a master teacher or head teacher of the respective school. This group focuses on evaluating teaching performance (i.e., peer-teacher evaluation). The second group comprises all of the students taught by the teacher in the respective year. Responding to a survey using a five-point Likert scale, students rate the teacher on various competencies and rate their level of satisfaction with their classes (i.e., student-class satisfaction). The third group includes the parents of these students, who also respond to a survey to rate their levels of satisfaction with their children's teachers and school. In the case of master teachers, the groups are similar, except that the principal, vice-principal, and head teacher do not necessarily need to act as evaluators. Finally, the principal and vice-principal of every school are evaluated by parents and teachers, but not by students. All evaluators have the opportunity to respond to a number of open-ended questions, as well (Choi & Park, 2016; Seo, 2012).

Teachers receive a score based on their evaluations by the different evaluators. The scores have significant consequences. Based on the evaluation scores, low-performing teachers are required to take additional hours of professional training in designated teacher training institutions. In contrast, teachers who receive high scores are rewarded with a 6- to 12-month sabbatical to concentrate on research (Choi & Park, 2016; Seo, 2012). Details of the criteria for determining the appropriate consequences based on evaluation scores by the different evaluators are presented in the discussion of components below. Figure 14.1 shows the multiple evaluators involved in the evaluation process and the outcomes based on the evaluation scores.

14.2.4 Influential Political Actors

The development of teacher evaluation policy in Korea has been influenced over the years by various political actors. The discussion of political influences and their effects on the development of the new teacher evaluation system, presented below, focuses on the major influences on the development of the new teacher evaluation system.

The influence of the media. One of the strongest influences on policy-making related to the new teacher evaluation system was the media. Although the public generally did not consider the traditional teacher evaluation system as an effective tool for improving student achievement and addressing the inequity of educational opportunities for all, it was not until the media released a report exposing that the public school system was not functional that the public demanded more school accountability. As mentioned previously, the media described this phenomenon as "school collapse" or "classroom collapse" (Whang et al., 2001), which referred to the failure of teachers to control the classroom, thereby leading to their failure to provide classroom lessons and maintain their authority. By focusing national attention on the serious problem of schools' inability to deliver the basic function of providing effective education to students, the media exerted a strong influence over effecting change in teacher evaluation policy (Kang, 2013).

The influence of the public's perceptions. Awareness of the "classroom collapse" caused the public to lose trust in the education system (Hwang, 2001; Mok, 2002; Park, 2006). Furthermore, the public criticized the traditional authoritarian and bureaucratic school system for its lack of response to the immediate needs of students and the community, thereby prioritizing efficiency over quality. Heated public discourse and debate on these issues were a powerful influence on school reform efforts, including the development of the new teacher evaluation system—Evaluation of Teacher Professional Development (Kang, 2013; Seo, 2012).

After the new teacher evaluation system was implemented, research was conducted to examine the perceptions of teachers, students, and parents on the effectiveness of the new teacher evaluation policy (Kim & Kim, 2012; Seo, 2012). The results showed that approximately 70% of teachers believed that the new system failed to help them identify their strengths and weaknesses, improve their teaching, or plan for improvement. On the other hand, the responses of students and parents were more positive. Approximately, 60% of students and 70% of parents believed that the new system had influenced teachers to make more effort in teaching. In 2019, a report submitted by the Ministry of Education provided data showing a decline in the participation rate of fellow teachers, students, and parents in the teacher evaluation system. This reduced participation could indicate that the evaluators lacked trust in the current system or that they considered the system to be ineffective (Han, 2020).

The influence of organized political forces. In Korea, various organized political actors have influenced policy decisions that affect all aspects of Korean society, including educational policy. The main organized political forces involved in the policy decisions related to the new teacher evaluation system were government actors, including the main administration, members of the majority conservative party, members of the liberal party, and the teachers' unions.

Governmental actors. As mentioned above, the government and the public had lost trust in the system due to the strong opinion that the traditional teacher evaluation system was not effective (Kang, 2013). Therefore, as the most powerful actor, the government responded to this problem by pursuing the development of a new system, with the support of the conservative legislators who were in the majority in Congress as well as members of the conservative party, Saenuri Dang. However, members of the liberal party, Minjoo Dang, were highly critical of the procedures for teacher assessment in the proposed new teacher evaluation system, in which they had had limited input. Therefore, they did not actively support the new policy in the initial stage of its development. The government's plan was to implement the new system in 2006; however, the lack of support delayed implementation until 2011 (Kang, 2013).

Teachers' unions. As a powerful political force, teachers' unions also played an important role in influencing the new teacher evaluation policy. The major teacher unions in Korea are the Korean Federation of Teachers' Association (KFTA) and the Korean Teachers and Education Workers Union (KTU). While the KFTA, primarily school administrators, supported the government's policy of developing a new teacher evaluation system, the KTU, composed of teachers, was critical of the policy. The new evaluation system received a strong backlash from teachers' unions, who expressed concern about its effectiveness, professional development, lack of consensus, and unreliable sources of evidence. Although students and parents responded more positively, a survey in 2005 reported that 88.5% of teachers agreed that the new teacher evaluation system would not improve student achievement, quality of education, and teacher quality (Kim & Kim, 2012; Seo, 2012). A questionnaire survey of 336 elementary school teachers in 2011 indicated that while some teachers believed that the teacher evaluation provided an opportunity to improve their teaching method, others raised issues related to student and parent evaluation. In particular, some teachers argued that students could make evaluations that were arbitrary, emotional, or mischievous and that some parents may use the evaluations as merely an opportunity to complain or just echoing their children's comments about their teachers (Choi, 2011).

In addition, in 2005, the KTU criticized the new system as being invalid and hastily developed, while the KFTA criticized the incentive policy proposed in the new system. Many big protests were held to express the unions' and teachers' objections to the new system, and in late 2006, the KTU made a dramatic statement of protest by occupying a public hearing and asserting that teachers would not follow the new system (Kim & Kim, 2012; Seo, 2012).

The strong negative opinions of the teachers' unions, composed of those who were most affected by the new teacher evaluation system, were primarily based on the following reasons (Seo, 2012). First, teachers criticized the new system for being too focused on requirements rather than professional development, focusing too much on accountability, and punishing less skillful teachers by forcing them to take professional training, thereby shaming and humiliating those teachers. Therefore, it is not surprising that, under the new system, teachers tend to give their colleagues high scores in peer evaluations, as giving low scores would have negative consequences for their peers. Furthermore, most teachers do not want to openly discuss their weaknesses or hear what they need to improve (Kim et al., 2011).

A second criticism has been the lack of consensus. The new evaluation system is based on a set of teachers' professional responsibilities in five areas: instructional design and planning, instructional implementation, assessment of student learning, individual student guidance, and fostering students' social competence. However, as the teachers are evaluated on their effectively engaging students in learning, the criticism has focused on the ambiguity of the term "effectively" and the lack of consensus on the performance indicators, standards of performance, or evaluation criteria. Moreover, the student and parent evaluations are considered by teachers to be arbitrary and highly subjective (Kim et al., 2011; Seo, 2012).

Third, teachers have criticized the new evaluation system for its unreliable sources of evidence. For example, despite the recommendations, peer reviewers make a single classroom observation and fewer than half of the parents observe even one class and are reluctant to provide feedback (Kim & Kim, 2012). Both evaluators and evaluated teachers believe that a single observation is insufficient. Consequently, due to the criticisms discussed above, teachers' unions consider the evaluations neither valid nor reliable (Seo, 2012).

14.2.5 Training of Evaluators and Key Components of the Current Teacher Evaluation System

In Korea, the new teacher evaluation system aims to improve student learning, educational equality, teachers' professional development, and parents' expectations of the quality of the educational system. Administrators, teachers, students, and parents are strongly encouraged to participate in the evaluation process; however, the participation rates have been decreasing in recent years. A 2011 interim report released by MEST showed that approximately 90% of teachers (343,725), 79% of students (4,191,548), and 46% of parents (3,045,765) participated in teacher evaluations in Korea in 2011 (Seo, 2012). In 2015, the participation rate of parents was 50%, but 35.21% in 2019 (Han, 2020). Discussed below are the procedures for training evaluators and the key components of the evaluation system including measures and consequences.[1]

14.2.6 Training of Evaluators

Evaluators are provided training at both the national and local levels on the standards of evaluation in order to secure reliability and to discuss the intent and goal of evaluation and the ethics and abilities required for carrying out evaluation duties. At the

[1] See https://files.eric.ed.gov/fulltext/EJ1128905.pdf for more detailed information on components as well as the overall framework of the current teacher evaluation system in South Korea including purpose, stakeholders, scope of evaluation, criteria and standards, and methods and instruments.

national level, training programs are conducted by the Korean Educational Development Institute (2010), which oversees research on educational policy for the South Korean Ministry of Education and Human Resource Development. Training consists of 15–20 h of lectures and workshops instructing evaluators on the system of school evaluation, including interpretation and application of indicators, evaluator ethics, and guide to writing reports. At the local level, half- to one-day training with lectures on the evaluation system and indicators is provided by the metropolitan/provincial offices of education for all evaluators within their respective districts. More recently, workshops have been held to allow trainees to practice indicator application (Korean Educational Development Institute, 2010).

14.2.7 Measures

Teachers' performance is assessed by the following measures: observations of classroom teaching and parents' and students' surveys.

Observations of classroom teaching. However the traditional teacher evaluation system relied exclusively on the principal's judgment of teacher performance, the new system involves multiple evaluations conducted by multiple evaluators. More specifically, classroom observations are conducted by principals, peer teachers, and parents. In the required peer review, at least three teachers and the school principal and vice-principal assess their colleagues' practices in more than one classroom observation. Parents are encouraged to observe their children's classrooms several times for teacher assessment before filling out their surveys (Choi & Park, 2016; Seo, 2012). Teachers are assessed based on criteria related to instruction and student guidance, described in more detail below, and receive a score based on their evaluations by the different evaluators.

Parents' and students' surveys. Parents respond to a survey to rate their levels of satisfaction with their children's teachers and school. Also, students in grades 4–12 are required to rate their level of satisfaction with their classes (i.e., student–class satisfaction). Using the same criteria as used for observations by administrators and peer teachers, as described below, parents and students evaluate the teachers (Kim et al., 2011; Seo, 2012). Evaluators score the teacher in a variety of competencies using a five-point Likert scale, with 1 being the lowest score and 5 the highest.

The criteria for evaluating teachers through observations and surveys focus on two areas—instruction and student guidance. As described by Choi and Park (2016), instruction consists of three elements: preparation, implementation, and assessment and utilization. Specific criteria for evaluating preparation include understanding the curriculum, showing evidence of efforts to improve teaching and learning methods and establishing teaching and learning strategies. Criteria used for evaluating implementation include teacher's attitudes, interactions between teachers and students, and instructional materials and activities. The third element of instruction, assessment and utilization, is based on other criteria, such as the assessment of student learning and the utilization of the results. The other focus of the criteria for teacher evaluation is student guidance, which consists of two elements: personal maturity and social maturity. Personal maturity is evaluated based on criteria such as developing students' strong characters and creativity and career guidance that considers students' aptitudes and strengths. Criteria used for evaluating social maturity include cultivating good habits and developing democratic citizenship (Choi & Park, 2016).

Consequences (utilization of evaluation results). Under the new teacher evaluation system, evaluation results are utilized in the following two ways: providing teachers who receive low evaluation scores with further teacher training through professional development opportunities and rewarding teachers who receive high evaluation scores with a sabbatical to concentrate on research, as illustrated in Fig. 14.1 (Seo, 2012).

As explained by Seo (2012), teachers who receive a score lower than 2.5 on their peer reviews and a score higher than 2.0 on the student surveys must take 60 h of professional training in designated teacher training institutions. Teachers who receive a score lower than 2.5 on the peer review and a score lower than 2.0 on the student surveys must take 210 h of professional training over six months. If these teachers fail to improve their scores the following year, they are removed from their classrooms for six months and must take 730 h of professional training at the National Training Institute of Education, Science, and Technology. In contrast, teachers who receive the highest scores can take a 6- to 12-month sabbatical to concentrate on research.

14.3 Assessment of the New (Current) Teacher Evaluation System

A 2011 interim report released by MEST showed that approximately 0.6% (2000) teachers received low evaluation scores and were required to take a specified number of hours of professional training. In contrast, about 0.2% (700) teachers who received high evaluation scores were rewarded with a sabbatical. The new evaluation system received a strong backlash from teachers, while students and parents responded more positively (Kim & Kim, 2012; Seo, 2012). The majority of teachers believed that the new system had little effect on their professional growth and failed to help them improve their teaching.

The new evaluation system received other criticisms as well. First, teachers criticized that the new system was too focused on requirements rather than support. Although the evaluation system emphasizes professional development, it actually focuses more on accountability. The system is considered by teachers to be shameful and humiliating in that it rewards effective teachers but punishes less skillful teachers by forcing them to take professional training.

Rather than serving as a disciplinary measure, professional development should be ongoing, as it is a natural and expected activity of teachers throughout their career and is key to school improvement (Huang, 2016). Therefore, as low evaluation scores would have negative consequences for their peers, such as forced professional development, it is not surprising that teachers in South Korea tend to give their colleagues high scores on peer evaluations. Furthermore, most teachers do not want to openly discuss their weaknesses or hear what they need to improve (Kim et al., 2011).

A second criticism of the evaluation system is the lack of consensus. The new evaluation system is based on a set of teacher professional responsibilities in five areas: instructional design and planning, instructional implementation, assessment of student learning, individual student guidance, and fostering students' social competence. Teachers are evaluated on several elements, including "engaging students in learning." To receive a high score, he or she must be evaluated as "effectively" engaging students. However, "effectively," as well as other terms used in the evaluation, can be considered ambiguous. Further, the evaluators have never established consensus on the performance indicators, standards of performance, or evaluation criteria. This ambiguity and lack of consensus have resulted in teacher distrust of the evaluation system.

Third, the student and parent evaluations are considered by many teachers to be arbitrary and highly subjective. Therefore, many teachers consider evaluations to be merely popularity contests and place little value on their results (Kim et al., 2011).

Fourth, teachers have criticized the new evaluation system for its unreliable sources of evidence. For example, although more observations are recommended, peer reviewers generally make a single classroom observation and fewer than half of the parents observe even one class and are reluctant to provide feedback (Kim & Kim, 2012). Both evaluators and evaluated teachers believe that a single observation is insufficient; consequently, teachers consider the evaluations neither valid nor reliable.

Despite the criticisms and the ineffective implementation of the new teacher evaluation system in many schools across the nation, some schools have successfully implemented the new evaluation system due to their unique efforts discussed below.

14.4 Successful Approaches to the New Teacher Evaluation System

The failure of the new teacher evaluation system in Korea is, in part, due to the fact that all schools are not the same and, therefore, what works in one school does not necessarily work in another school (McLaughlin, 1976, 1990; Thorn & Harris, 2013). In light of criticisms of the new teacher evaluation system, some schools in Korea have demonstrated new approaches to the system, which have produced better results. Three school cases (elementary, middle, and high school) were selected from research presented by Seo (2012), based on her review of a 2012 report released by MEST. This report is considered a reliable source due to the comprehensive manner in which it was produced. As required by MEST, the teacher evaluation committee in each school must aggregate all teacher reports and produce a school report to turn into local educational agencies (LEAs). This final school report is open to the public for teachers and parents to review. Finally, after collating the final school reports sent by LEAs, MEST releases the interim report every year to the public. This report, as well as Seo's (2012) research, highlights the three cases described below as examples of schools that have modified the national teacher evaluation system for successful implementation at the individual local level.

14.4.1 Hangang Middle School

As described by Seo (2012), Hangang Middle School, located in Seoul's Yongsan District (Yongsan-gu), is one of the most ethnically diverse regions in South Korea. The student population in the three grades totals 538, with a faculty of 46 teachers. In the second year after implementing the new evaluation system, teachers at Hangang Middle School found that the standardized evaluation questions and rubrics in the new evaluation system were not useful for identifying teachers' strengths and weaknesses or improving their practices. To overcome these weaknesses of the new system, the teachers developed a school-level teacher evaluation system while maintaining the framework of the national system. The teachers added questions focused on student learning, including three open-ended questions for the peer review and two additional questions using a five-point scale (e.g., "Was the observed lesson effective in helping students learn key concepts in the subject matter?"). Post-observation meetings were held to allow peer reviewers and the teacher the opportunity to review the key concepts presented in the observed lesson and to discuss whether students learned those concepts and whether there was evidence illustrating that learning. The teachers discussed the survey results with both students and parents and reached a consensus that clarified what was expected of teachers and what to assess in teacher evaluation. All participants agreed that good teaching involved engaging all students in learning, helping students develop a deep understanding of the content, effectively communicating with students, caring for individual students, and helping all students succeed in school. Through this new approach of the school-based teacher evaluation system, teachers were provided constructive feedback that was useful for improving their teaching. Furthermore, the teachers reported that this system was quite successful in promoting their professional development (MEST, 2012; Seo, 2012). Hangang Middle School's successful adaptation of Korea's national evaluation system illustrates the theoretical view of Thorn and Harris (2013) that new roles and relationships are reorganizing public education. Specifically, according to their perspective, individual school leaders and teachers can work together to modify the new evaluation system for successful implementation in their school, to help them respond to the pressures of accountability required by the new system, and to improve their professional development (Seo, 2012).

14.4.2 Namsan High School

Namsan High School, located in Gyeonggido, has 821 students in the three grades, with a faculty of 88 teachers. As explained by Seo (2012), the teachers recognized problems with the new national evaluation system and therefore developed a school-level peer review system that encourages collaboration among the teachers. Teachers in the same department form teams of four or five to diagnose the needs of their individual students and to plan appropriate lessons together to address those needs. As performing well on the college entrance exam in the 12th grade is a top priority for 12th-grade students at Namsun High School, the teachers focus on promoting students' academic achievement. The teachers modified the national system by requiring each team member to conduct a lesson at least twice each semester, while other team members observe. After the observation, the team meets and evaluates the lesson in order to identify the strengths and weaknesses in the teacher's practice and provide suggestions as to what the teacher might do to improve. At the end of the school year, the peer reviewers submit the final scores for each team member. Seo (2012) reported that this collaborative approach significantly improved student achievement and a high evaluation rating for the school, which led to receiving monetary incentives from the Gyeonggido Office of Education. Namsan High School's approach demonstrates Thorn and Harris's (2013) view that collaboration is essential for success. By focusing on collaboration and the collective responsibility of all parties involved (teachers, parents, and the students themselves), teachers at Namsan High School were successful in modifying the new teacher evaluation system for their professional development and student achievement.

14.4.3 Sejong Elementary School

The third example of successful adaptation of the national teacher evaluation system described by Seo (2012) is Sejong Elementary School, located in a densely populated district on the north bank of the Han River, to the eastern end of Seoul. The number of students in the six grades totals 365, and the school's faculty includes 19 teachers. After implementing the new teacher evaluation system at their school, the teachers found that fewer than half of the parents participated in the teacher evaluation process. Furthermore, even when the parents participated in the teacher evaluation process, their evaluations tended to rely on what their children said about their teachers. Moreover, teachers could not understand the reasons behind their scores and what those scores meant. Therefore, teachers felt that the manner in which parents participated should be changed. Through online message boards, e-mails, text messages, and bimonthly teacher–parent meetings, the teachers began communicating regularly with parents to discuss their children's learning styles and ways in which parents could more effectively help their child with homework. As a result, teachers developed a better understanding of individual students, which enabled them to provide the appropriate support. Another result of implementing the necessary changes was that approximately 70% of parents participated in the teacher evaluation process (MEST, 2012; Seo, 2012), not only doing a better job of rating teacher performance but also providing more helpful feedback to the teachers. For example, some parents suggested that after exams teachers should review questions that students had answered incorrectly. The success at Sejong Elementary School demonstrates how teachers and parents can build a shared vision and shared values through collaboration to foster student learning and develop competencies for teacher evaluation (Seo, 2012). The case of Sejong Elementary School also underscores Thorn and Harris's (2013) theoretical perspective on the importance of teachers communicating regularly with parents and all actors working together to successfully implement the national system at the local level as well as to help teachers respond to the pressures of accountability.

Focusing on improving parent participation and parent–teacher communication, teachers at Sejong Elementary School were successful in modifying Korea's new teacher evaluation system for their professional development and student achievement.

14.5 Discussion

Korea's national teacher evaluation system has evolved over the decades, culminating in the current system—Evaluation of Teacher Professional Development—shaped by various factors including legislative actions, policy directives, and political influences. Based on their evaluations by various evaluators, teachers receive scores that have significant consequences, aiming to improve teacher quality and, in turn, student achievement.

Many concerns and criticisms have resulted in distrust of the evaluation system (Kim et al., 2011; Seo, 2012). One of the major concerns is the overall effectiveness of the system. Choi and Park (2016) determined that Korea's previous and current evaluation systems failed to meet all of the criteria for an effective teacher evaluation system. To evaluate all aspects of a teacher evaluation system and to determine what future modifications are needed for improvement, Darling-Hammond's (2012) criteria can serve as a comprehensive assessment tool:

(1) The teacher evaluation should be based on professional teaching standards;
(2) Evaluations should include multifaceted evidence of teacher practice, student learning, and professional contributions;
(3) Evaluators should be knowledgeable about instruction and well trained in the evaluation system;
(4) Evaluation should be accompanied by useful feedback and connected to professional development opportunities;
(5) The evaluation system should value and encourage teacher collaboration;
(6) Expert teachers should be part of the assistance and review process;

(7) Panels of teachers and administrators should oversee the evaluation process. (Darling-Hammond, 2012, p. 38).

According to Choi and Park's (2016) assessment, the current national system in Korea meets only two of Darling-Hammond's seven proposed criteria: panels of teachers and administrators should oversee the evaluation process; and evaluation should be accompanied by useful feedback and connected to professional development opportunities. Therefore, a more serious effort to address the unmet criteria is required.

An effective teacher evaluation system requires commitment and support from all involved. It is imperative that schools build a consensus on understanding standards of performance and evaluation criteria among teachers, principals, students, and parents, all of whom have the responsibility of fair and effective evaluation of teachers. Therefore, to ensure that all actors involved in teacher evaluation in Korea have the same understanding of performance standards and evaluation criteria, evaluation training should be required at the beginning of each school year to review all evaluation guidelines for all evaluators.

Validation studies should be developed to reduce the impact of measurement error. A number of studies have explored important research questions about the validity of teacher performance assessment instruments, especially when using classroom observations. For example, the research questions guiding a study by McEachin et al. (2018) were (1) to what extent does content expertise of the teacher evaluation rater influence evaluation scores and (2) to what extent are teacher observation scores valid predictors of the effectiveness of teachers, as assessed by their contributions to student performance (or value-added scores). Based on their findings, McEachin et al. (2018) recommended the following:

1. Consider setting higher standards for rater certification (specifically by requiring raters in training to have their scores align with those of master raters at a higher frequency).
2. Research has shown that raters tend to change their approach to scoring over time—a phenomenon known as rater drift. To mitigate this, consider using more frequent post-certification calibration and validation exercises during a rating period.
3. Keep in mind that high-quality evaluation and feedback require many observers with different backgrounds to rate many lessons.
4. Consider collecting additional sources of evidence that support claims about the quality of teacher practice, i.e., a multiple-measure system.

Such future studies could not only reveal validity issues related to the Korean teacher evaluation system but also provide valuable insight into developing a more effective assessment of teacher performance.

Furthermore, Darling-Hammond and McLaughlin (1995) asserted that staff development means allowing teachers critical reflection on their practice, which can drive them to develop new knowledge and beliefs about content, pedagogy, and learners. Teachers in Korea need to be provided the opportunity and adequate time to reflect on their evaluation feedback each semester so that they can make the necessary adjustments to improve their performance.

Although the consequences of the present Korean system are incentives for teachers to improve in performance, receiving a low evaluation score and considering the required hours of professional development as punishment inflicts shame and humiliation. To avoid this perceived negative consequence, other ways to improve teacher quality should be investigated. For example, low-performing teachers could be mentored by senior teachers who have received high evaluation scores. Also, as demonstrated in one of the successful case studies presented in this paper, teachers can organize teams to encourage collaboration among teachers in the same department. The teachers evaluate their peers' performance and identify the strengths and weaknesses in the teacher's practice, providing constructive criticism and positive suggestions. In sum, successful evaluation can be achieved through collaboration and the collective responsibility of all parties involved (teachers, parents, and the students themselves).

Most important, as local schools have their individual differences and needs. In other words, what may be appropriate for one school may be inappropriate for another school. As McLaughlin (1990) asserted, the success of educational policy implementation at the local level depends on mutual adaptation with consideration of the fundamental and consequential differences of individual schools, rather than uniform implementation enforced by federal policy. Thus, each school in Korea has the responsibility of adapting the standard evaluation criteria that best fits their school by relying on current educational research findings (e.g., Choi & Park, 2016; Darling-Hammond, 2012; Seo, 2012; Thorn & Harris, 2013) and following the examples of the three case studies of successful adaptation and implementation discussed in this chapter.

The main concern of Korea's Ministry of Education is that every school makes whatever changes are necessary to improve teacher quality in order to improve student achievement. The school administration, faculty, parents, and students are encouraged to adapt the evaluation system so that it is effective and appropriate at the local level. Rather than rigid enforcement of the national evaluation criteria, individual schools must focus on building an environment in which principals, teachers, parents, and students collaborate for successful local mutual adaptation of the national evaluation policy. One further possible action that the Ministry of Education can take to promote such collaboration and adaptation is to offer workshops, conferences, and local discussions that provide clear guidance and support to local actors. Thus, for a teacher evaluation system to be successful in each school, accountability, necessary modification, and mutual adaptation are required (Yoo, 2018).

In Korea, the continuing efforts and input of all actors at both the national and local levels are needed to develop and implement the most effective national teacher evaluation system whose goals are to achieve equitable distribution of highly effective teachers throughout the nation, thereby ensuring that every student has equal access to the best education. Thus, the evolution of the teacher evaluation system in Korea continues.

References

Choi, H., & Park, J. (2016). An analysis of critical issues in Korean teacher evaluation systems. *Center for Educational Policy Studies Journal*, *6*(2), 151–171. https://files.eric.ed.gov/fulltext/EJ1128905.pdf

Choi, J. (2011). *A study on the teachers' perception about 'the teacher evaluation program for professional development'* (Unpublished master's thesis, Seoul National University of Education, Seoul, South Korea). https://academic.naver.com/article.naver?doc_id=80328007

Coolahan, J., Santiago, P., Phair, R., & Ninomiya, A. (2004). *Attracting, developing, and retaining effective teachers–country note: Korea.* OECD Education and Training Policy Division.

Darling-Hammond, L. (2012). *Creating a comprehensive system for evaluating and supporting effective teaching*. Stanford Center for Opportunity Policy in Education.

Darling-Hammond, L., & Mclaughlin, M. W. (1995). Policies that support professional development in an era of reform. *Phi Delta Kappan*, *76*(8), 597–604.

Goldhaber, D. (2002). The mystery of good teaching. *Education next, 2*(1), 50–55.

Han, M. (2020, May 24). Teacher evaluation participation rate for 5 years. *Seoul Economic Daily*. https://www.sedaily.com/NewsVIew/1Z2VJKLU3Q

Huang, B. (2016). Transformation and framework of teacher professional development in Taiwan. *Policy Futures in Education, 14*(7), 926–942.

Hwang, K. (2001). Patterns of on-line discourses about the school failure. *Theory and Research in Citizenship Education, 33*, 407–438.

Jeon, J. (2001). Kyo-won pyung-ga-wa kyo-won sung-kwa-keup [Teacher evaluation and bonuses]. *Kyo-Yook Yee-Ron-Kaw Sil-Cheon [educational Theory and Practice], 11*(2), 143–186.

Kang, N. (2013). Teacher evaluation policy development in South Korea. In M. Akiba (Ed.), *Teacher reforms around the world: Implementations and outcomes* (Vol. 19, pp. 147–177). International Perspectives on Education and Society.

Kang, S. (2001, September 26). Where are teachers standing in Korean society? *Eduhope*. http://news.eduhope.net/9590

Kim, K. S., Jeon, J. S., & Ahn, B. C. (2011). *Developing a teacher evaluation model for professional development*. Korean Educational Development Institute.

Kim, K. S., & Kim, E. K. (2012). *The results of 2011 teacher evaluations*. Korean Educational Development Institute.

Korean Educational Development Institute. (2010). *Country background report for Korea*. https://www.oecd.org/education/school/49363138.pdf

McEachin, A., Schweig, J., Perera, R., & Opper, I. M. (2018). *Validation study of the TNTP core teaching rubric (RR-2623-NTP).* RAND Corporation. https://www.rand.org/pubs/research_reports/RR2623.html

McLaughlin, M. W. (1976). Implementation as mutual adaptation: Change in classroom organization. *Teachers College Record, 77*(3), 339–351.

McLaughlin, M. W. (1990). The RAND change agent study revisited: Macro perspectives and micro realities. *Educational Researcher, 19*(9), 11–16.

Ministry of Education, Science, and Technology. (2006). *Plans for improving the teacher evaluation system.* Seoul, South Korea.

Ministry of Education, Science, and Technology. (2012). *Plans for improving the teacher evaluation system.* Seoul, South Korea. www.mest.go.kr/web/1110/ko/board/view.do?bbsId=149&boardSeq=27378

Mok, Y. (2002). Kyo-yook jun-tong yuk-hak goo-do byun-wha-wa kyo-sil boong-kwae kwan-lyun sung yun-goo [A study of relation between changes of mechanism in educational tradition and the classroom collapse]. *The Journal of Educational Idea, 11*, 35–51.

Organization for Economic Co-operation and Development [OECD]. (2009a). *Creating effective teaching and learning environments: First results from TALIS*. Centre for Educational Research and Innovation, Organization for Economic Co-Operation and Development.

Organization for Economic Co-operation and Development [OECD]. (2009b). *Equity in education: Students with disabilities, learning difficulties and disadvantages*. Centre for Educational Research and Innovation, Organization for Economic Co-Operation and Development.

Park, K. (2006). Kong-kyo yook jung-sang-hwa wee-han kyo-yook jung-chaek kyul-jung chejecham-yuh-ja yuk-hal jo-myung [The roles of decision-makers in educational policy for the revitalization of public education]. *The Journal of Research in Education, 26*, 173–204.

Pham, L. D., Nguyen, T. D., & Springer, M. G. (2021). Teacher merit pay: A meta-analysis. *American Educational Research Journal, 58*(3), 527–566. https://doi.org/10.3102/0002831220905580

Polanyi, M. (1958). *Personal knowledge: Towards a post-critical philosophy*. Routledge & Kagan Paul.

Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, schools, and academic achievement. *Econometrica, 73*(2), 417–458.

Rockoff, J. E. (2004). The impact of individual teachers on student achievement. *American Economic Review, Papers & Proceedings, 94*(2), 247–252.

Seo, K. (2012). Lessons from Korea. *Association for Supervision and Curriculum Development, 70*, 75–78. https://sites.miis.edu/comparativeeducation/files/2013/01/Lessons-from-Korea.pdf

Thorn, C., & Harris, D. N. (2013). The accidental revolution: Teacher accountability, value-added, and the shifting balance of power in the American school system. In D. Anagnostopoulos, S. A. Rutledge, & R. Jacobsen (Eds.), *The infrastructure of accountability: Data use and the transformation of American education* (pp. 57–74). Harvard Education Press.

Whang, K., Yang, E., Jun, Y., & Hug, H. (2001). "Kyo-sil boong-gwae"-ro bull-li-neun kyo-sa-hak-saeng gal-deung hyun-san ee-hae wee-han jill-juk yun-goo [A qualitative study about tension between teacher and students in the phenomenon called, "classroom collapse."]. *The Journal of Korean Education*, *28*(2), 247–276.

Yoo, I. (2009*). Kyo-won neung-ryuk pyeon-ga je-do jung-chak sa-le boon-suk* [*Analysis of teacher competence evaluation policy case*]. Ministry of Education, Science and Technology.

Yoo, J. (2018). Evaluating the new teacher evaluation system in South Korea: Case studies of successful implementation, adaptation, and transformation of mandated policy. *Policy Futures in Education, 16*(3), 277–290. https://doi.org/10.1177/1478210317751274

Chapter 15
The Loss of Teacher Appraisal in New Zealand: A Theory-of-action Perspective

Claire Sinnema

Abstract The requirement for teachers to be appraised annually was recently removed from the New Zealand education system. In response to claims that appraisal was burdensome, created too much workload and had too little impact, it was removed and replaced by a process called a professional growth cycle. In this chapter, I first introduce the former appraisal system and its key features. I then examine the shift to the new process using a theory-of-action approach to describe both the former and current systems. The theory-of-action approach draws attention to the constraints driving the practice of policy makers for both systems and allows a critique of the move to abandon rather than improve appraisal in ways consistent with double-loop learning in a system that learns. While the new system is still recent and the consequences of it are uncertain, I argue that much of what the new system emphasizes could have been applied to improving the pre-existing appraisal process. There was potential, I suggest, to retain an accountability mechanism in appraisal alongside moves to increase trust and reduce burden. Furthermore, such integration of accountability alongside improvement purposes is particularly important in systems seeking to address issues of educational inequity.

15.1 A Theory-of-Action Account of the Appraisal Policy Context

Equity issues are of concern in the education system in New Zealand, of more than 71,000 teachers—approximately 41,000 teaching in the primary sector and 30,000 in the secondary sector. Most work in state schools, while just under 9000 work in state integrated schools (typically with religious special character) and just under 2000 in special schools. Most of the teachers in both sectors are female. A notable proportion of teachers, more than 10%, are over the age of 65. While there is a national education system, and schools are required to implement the national curriculum,

C. Sinnema (✉)
The University of Auckland, Auckland, New Zealand
e-mail: c.sinnema@auckland.ac.nz

J. Manzi et al. (eds.), *Teacher Evaluation Around the World*,
Teacher Education, Learning Innovation and Accountability,
https://doi.org/10.1007/978-3-031-13639-9_15

there have been high levels of autonomy in schools, including for the design of curricula at the local level (Sinnema, 2015). Schools are self-governing with Boards of Trustees at the local level responsible for the governance of individual schools. In total, there are just over 2500 state and state integrated schools in New Zealand (1943 primary, 178 composite, 378 secondary and 37 specialist). While there are a number of private schools, national curriculum expectations and the appraisal requirements described here do not and have not applied to those schools. Relevant to the focus on teacher appraisal discussed here is the attention in the New Zealand system to evidence informed practice (Lai & Sinnema, 2022) and collaborative approaches to educational improvement (Sinnema et al., 2021).

15.1.1 The Introduction of Teacher Appraisal in the New Zealand Education Policy Context: The 1980s

For thirty years, between 1989 and 2019, schools in New Zealand held accountability for managing the performance of teachers. That accountability was instituted as part of the 1989 Tomorrow's Schools reform (Government of New Zealand, 1989c) together with the Education Act (Government of New Zealand, 1989a). The New Zealand education reforms brought a fundamental change to the governance of schools. The turn to decentralized decision-making saw Boards of Trustees established for each school, comprised of parent elected, staff and student members. Responsibility was placed on Boards for all aspects of governance including curriculum, personnel, financial and property matters. The State Sector Amendment (Government of New Zealand, 1989b) set out for Boards their responsibility, as Piggot-Irvine (2000) describes, relating to the enhancement of development (Section 77A), maintenance of standards of integrity and conduct (77A) and the assessment of the performance of teachers (77C). Despite all of these obligations, and efforts in some contexts to meet them, no specific prescription, or guidelines, for the implementation of appraisal existed prior to 1996.

15.1.2 The Move to Performance Management Linked to Professional Standards: The 1990s Onwards

Nearly a decade later, the requirements that Boards of Trustees develop and implement performance management policies and conduct performance appraisal of their teachers were supported by the introduction of Guidelines for Performance Management in Schools (The New Zealand Ministry of Education, 1997, 1998a, 1998b). While described as guidelines, they had mandatory status. The guidelines clarified the expectations on Boards to implement personnel policies that complied with good employer principles and that ensure teachers to provide an education that fully meets their students' needs. They were responsible for teachers' performance expectations, performance appraisal, reward systems, professional development and, where necessary, disciplinary/competency procedures. The purposes of appraisal were clearly laid out:

1. [Appraisal should] "provide a positive framework for improving the quality of teaching (and therefore learning) in New Zealand schools."
2. [Appraisal should] fulfill both an accountability (summative) and development (formative) function.
3. Minimum standards of accountability and quality assurance should be set out, that are flexible enough for Boards of Trustees to establish appraisal systems appropriate to their school and community.
4. Appraisal should be integrated into schools' planning and organization for professional development.
5. School-wide goals and objectives should be linked with the professional goals of individuals.
   (The New Zealand Ministry of Education, 1998a, 1998b, p. 1).

The aspects of teaching performance that were required to be appraised included (1) teaching responsibilities (planning and preparation, teaching techniques, classroom management, classroom environment, curriculum knowledge and students assessment); (2) school-wide responsibilities (e.g., contribution to curriculum leadership, school-wide planning; school goals; the effective operation of the school as a whole; pastoral activities; and student counseling and community relationships); and (3) management responsibilities (e.g., planning; decision-making; reporting; professional leadership; and resource management). The elements of the process that were set out in the performance management guidelines included the following:

> Identification of an appraiser, in consultation with the teacher concerned[1]
>
> Development of a written statement of performance expectations,[2] in consultation with each teacher
>
> Identification and written specification of one or more development objectives to be achieved during the period for which the performance expectations apply
>
> For each development objective, the identification and written specification of the assistance or support to be provided
>
> Observation of teaching
>
> Self-appraisal by the teacher
>
> Opportunity for the teacher to discuss their achievement of the performance expectations and the development objective(s) with their appraiser
>
> An appraisal report prepared and discussed in consultation with the teacher
>
> (The New Zealand Ministry of Education, 1997, p. 5)

[1] Typically, a colleague senior to or more experienced than the teacher would be appointed as the appraiser, and in some cases, external appraisers from outside of the school were appointed.

[2] Schools had autonomy to decide on the performance expectations referred to here.

Policies and processes were required to be guided by a set of seven principles. These required appraisal to: (1) have a professional development orientation, (2) be appropriate to individual teachers, the school and the wider community, (3) be developed in a consultative manner with teachers, (4) be open and transparent, (5) be part of an integrated performance management system operating within the school, (6) be timely and helpful to the individual teacher and (7) give consideration to matters of confidentiality, including the provisions of the Privacy Act and the Official Information Act (The New Zealand Ministry of Education, 1997, pp. 4–8).

The dual focus on development and accountability was a feature of the appraisal system at that time. While the stated primary purpose of appraisal requirements was to provide a positive framework for improving the quality of teaching (and therefore learning) in New Zealand schools, the guidelines also set minimum standards of accountability and quality assurance; the guidelines were, however, flexible enough for Board of Trustees (each school in New Zealand's self-governing school system is governed by a Board of Trustees; a crown entity responsible for schools' performance and legal obligations) to establish systems appropriate to their particular school and community. The accountability aspect was evident at both the local level (in that senior managers and the Board were required to be informed of appraisal outcomes) and the national level (subsequent policies created a stronger link between appraisal and remuneration for all teachers across the country). While the salary scale for teachers is established nationally, decisions about teachers meeting their accountabilities in order to progress through the scale are made locally. Because the same processes were usually used for school leaders to establish if teachers met the New Zealand Teaching Council's Registered Teacher Criteria (necessary for gaining or continuing teacher registration), the accountability aspect was also prominent.

Soon after the introduction of the mandatory guidelines, professional standards for primary teachers were introduced into the New Zealand system (The New Zealand Ministry of Education, 1998a, 1998b). The professional standards were linked to the Primary Teachers' Collective Employment Agreement, and it became a requirement to assess these as part of each teacher's appraisal. The seven professional standards' dimensions at that time were professional knowledge; teaching techniques; motivation of students; classroom management; communication; support for and cooperation with colleagues; and contribution to wider school activities. The standards were accompanied by indicators (24 for a fully registered teacher). Guidance was developed in response to requests from schools for a practical tool to integrate the Interim Professional Standards into performance management systems. The objective of performance management in schools was, according to this guidance, to improve learning outcomes for students by improving the quality of teaching and leadership. That developmental purpose was in conjunction with the accountability function of the professional standards, since they needed to be met in order to progress up the salary scale. At the same time, in the secondary sector, an attestation process was established, which included teachers' completion of appraisal in a set of criteria to be met as part of attestation, and also linked to remuneration decisions.

15.1.3 A Turn to Embedding Teaching as Inquiry into Appraisal: 2007 Onwards

In 2007, a new national curriculum was introduced in New Zealand. That curriculum included a model of effective pedagogy to support the realization of curriculum aspirations, focused on a model of Teaching as Inquiry. While not compulsory, the model was prominent in the 2007 curriculum reform, and included in the guidance (rather than prescription) part of the curriculum statement, and was central to a range of other professional learning and research and development initiatives. The Teaching as Inquiry model had its origins in a Best Evidence Synthesis of effective pedagogical approaches to support learning in the Social Sciences (Aitken & Sinnema, 2008). In Teaching as Inquiry, educators engage in three kinds of inquiry as they seek to achieve curriculum goals—focusing inquiry, teaching inquiry and learning inquiry (Sinnema & Aitken, 2011, 2013, 2019). In the focusing inquiry, teachers pay careful attention to prioritizing what matters most for students given the curriculum requirements, community expectations and, most importantly, the learning needs, interests and experiences of the learner. In the teaching inquiry, they give close attention to two sources of evidence—outcomes-linked research evidence and practitioner experience. They are asked to use that evidence to inform decisions about what teaching strategies are most likely to work and are therefore worth trying. The Teaching as Inquiry model encourages teachers to view research evidence as the basis for explaining findings about the impact of their own practice on their students' learning and as sources of better-informed conjectures about what might enhance learning for students in their classrooms. In the learning inquiry, it requires consideration of the impact of teaching actions on student outcomes and experience, as well as inquiry into the relationship between the teaching and those outcomes.

The introduction of Teaching as Inquiry was well received in many schools, and the national evaluation agency (The Education Review Office) found in a 2012 national evaluation of the implementation of Teaching as Inquiry that 58% of schools had processes in place that were either highly or somewhat supportive of Teaching as Inquiry (Education Review Office, 2012). While Teaching as Inquiry was not compulsory, it gained status through its inclusion alongside compulsory aspects of the national curriculum. Many schools began to integrate Teaching as Inquiry with teacher appraisal. But the quality of the approaches taken to inquiry was variable. The Education Review Office (2012) reported that "teachers and leaders were stronger at the focusing inquiry phase (identifying which students need help), than they were at planning how to respond to them (teaching inquiry) and evaluating how well programmes impact on learners (learning inquiry). These latter stages require a level of problem-solving and evaluation that challenge many teachers" (p. 1). They highlighted the need for development of leaders' competencies for leading inquiry in ways that promote improvement in teaching and learning. While they noted evidence of "clear benefits for students and teachers when inquiry happens well" (p. 2), they also highlighted the demand for timely responses to students' needs and strengths and good feedback loops for when teachers observe, respond and evaluate in real time.

15.1.4 *The Tendency for Compliance-Oriented Appraisal with Limited Impact on Improvement*

The issues identified above with regard to the quality of Teaching as Inquiry approach mirrored earlier (Sinnema, 2005) and later (The New Zealand Education Review Office, 2014) studies related particularly to teacher appraisal. In particular, this work highlighted weaknesses in the extent to which appraisal supported the development/improvement function of appraisal.

Sinnema (2005), for example, in a series of studies, found only limited attention was given, in critical elements of teacher appraisal, to student learning. Appraisal was, therefore, lacking in terms of its conduciveness to improving teaching and learning. In school's appraisal policies and supporting documentation, indicators used by schools to evaluate teachers seldom focused directly on student learning. Appraisal discussions, similarly, typically focused on teacher practices without exploring connections between those practices and the impact they have on student learning. Teachers rarely reported discussions that included talk specifically about student learning, and none reported reference to student learning data.

A study of appraisal goals highlighted their considerable influence on the content and scope of appraisals which was problematic given the vast majority of goals focused on teaching practice and very few (5%) referred to student outcomes. There were also issues with regard to the specificity of goals in terms of signaling which learners and which aspect of learning were the target of the goal, and in terms of the aspirations, goals referred to which were often generic (e.g., to improve "learning" rather than to improve a more specifically detailed aspect of learning); the extent of improvement teachers were aiming for was either not dealt with at all or ambiguous. Overall, goals tended to be vague, rather than specific, and were not explicitly challenging. A model of appraisal called "appraisal for learning" was argued for, an approach that focuses on teacher learning about student learning and emphasizing both accountability for quality of inquiry into the impact of teaching on learning, but with a strong leaning toward the formative and developmental goals of appraisal at the time (The New Zealand Ministry of Education, n.d.-b).

Nearly ten years later, the Education Review Office (ERO) investigated approaches to teacher appraisal in New Zealand schools. During their scheduled reviews of schools, and through online surveys responded to by 173 schools, they found examples of some schools taking robust and effective approaches:

> schools in this study with highly robust appraisal processes balanced a professional accountability focus with a strong desire to make improvements for their students. They looked deeply into student achievement results to determine the impacts of changes in teaching practice and to decide what aspects of their teaching they needed to improve. Necessary teaching improvements identified through *Teaching as Inquiry* often contributed to their appraisal goals. Teachers recognised the relationship between effective appraisal, strengthened professional practice and the ongoing processes used in the school to identify and support improvement. High quality teacher appraisal was implemented as part of the planning and reporting cycle in the most successful schools. It was linked to the goals of the strategic plan, to the annual plan, to the principal's performance management system, and to

> decisions about teacher professional learning and development (PLD). (The New Zealand Education Review Office, 2014, p. 1)

However, in the majority of schools, they found appraisal did not contribute sufficiently to improving teacher capability and student outcomes. Most of the schools reviewed had compliant appraisal systems that included all the accountability aspects required, but there was limited evidence of their appraisal systems being integral to overall school improvement efforts.

15.1.5 The System Move to Auditing of Appraisal by External Review Agency: 2015–2019

In 2015, there was a turn in appraisal policy and practice in New Zealand that increased the emphasis on accountability, perhaps not surprising given the issues reported above related to the robustness and efficacy of the process across the system. New Zealand's Education Council (the professional body for teachers in New Zealand that sets standards, registers teachers, provides professional leadership, approves initial teacher education programs and consults on key policy developments and the like) contracted the Education Review Office for three years to provide an independent audit of appraisal across the various education settings.

The purpose of this audit process was to monitor whether two requirements in the Education Act were met: 1) that appraisals support the issue and renewal of practicing certificates and 2) that appraisals are of a "reasonable and consistent" standard. It was to provide an "across the system" picture of the quality of appraisal systems and was important for developing an understanding of what was happening in the profession. It also provided the public with additional, independent assurance (The New Zealand Education Council, 2018).

During their routine evaluation visits to schools, ERO used two types of indicators for good practice (see below) to audit the appraisal process for a at least 10% of the practicing certificates issued or renewed each year—this involved 4000 individual audits. The indicators related to both the individual level (see Table 15.1) and the system level (see Table 15.2). At the individual teacher level, the indicators focused on whether the endorsements in the sample audited (those endorsed in the previous 12 months) were based on "meaningful" appraisal, was evidence linked to the professional teaching criteria, and was the evidence that each criteria were being met necessary and sufficient.

At the system level, the audits focused on whether appraisals by professional leaders achieved a "reasonable and consistent" standard overall.

The audit process functioned, according to ERO (The New Zealand Education Council, 2018), as an incentive for promoting improved practice—in other words, it was a catalyst for improved processes. While the audits did not find universally high-quality practices, they did note improvements from one year of the audit to the next:

Table 15.1 Individual level indicators

Individual level indicators
• Personalized appraisal process
• Targeted observation of teaching and links between teaching practices and student/ākonga learning. Appraisal includes reflection about practice and outcomes for learners
• Teaching as inquiry
• Range of robust information used, including perspectives of students/ākonga and parents
• High-quality feedback about teaching practice and next steps provided
• Appraisal goals linked to ākonga learning/outcomes/wellbeing and the school's/service's strategic goals
• Appraisal goals are specific and can be verified by objective measures or indicators
• Appraisal identifies support and professional learning and development needed
• Opportunities for data-based discussion between teachers and leaders about student/ākonga learning and its relationship to teaching
• Endorsement of leaders' performance based on appropriate appraisal using professional teaching criteria

Table 15.2 System level indicators

System level indicators
• Senior leader responsible for both completion and quality of appraisals
• Senior leader who makes final endorsement decision is assured of the quality and breadth of appraisal process and evidence
• Processes are well documented to support the teacher's application for the practicing certificate
• Clear comprehensive procedures guide appraisal practice, including using the PTCs. These might include, for example, developing worthwhile and specific goals, indicators, robust evidence including achievement information, classroom observations, self-appraisal and the final report
• Effective processes are used for induction and mentoring of teachers to be recommended for the issue of a full practicing certificate and for those working toward full certification
• Templates and observation schedules provide guidance about goals, process, evidence and observation of teaching
• Time is allocated for goal setting, appraisal observations and discussions
• PLD on effective appraisal processes and evidence, using PTC, providing constructive feedback, and coaching, promoting consistent understanding of expectations for teaching
• Board is assured about teacher status—certification/endorsement and completion of appraisal endorsement and appraisal procedures and practices reviewed and improved regularly

> There was a small but steadily improving trend evident during the second year in the quality of appraisal that supported the endorsements made by professional leaders for the issue and renewal of practicing certificates. In the 2016-17 year, the percentage of issues of a Full Practicing Certificate that were judged as based on a satisfactory process increased from 77% to 83% overall. Satisfactory renewals of practising certificates improved from 65% to 74%...As was the case in the first year, many schools and services had revised and improved their overall appraisal systems either just prior to ERO's audit or in the year leading up to it.

> Therefore, where a teacher may not have had evidence of regular appraisal (incorporating the practicing teacher criteria during all the previous three years), the current cycle was more likely to be both compliant with requirements and more meaningfully focused on improving teaching. (p. 3)

15.1.6 New Professional Standards and a Focus on Evidence: 2018

In 2017, the professional standards for teachers in New Zealand were updated (Education Council of Aotearoa New Zealand, 2017). These most recent *Standards for the Teaching Profession* are made up of six standards (Table 15.3).

For each standard, there are a set of elaborations that provide additional detail depth and context to the standards themselves. The elaborations are intended to support teachers to recognize and develop the quality of their own and others' practices (see Table 15.4).

These standards continued to signal both development and accountability purposes of appraisal (the context in which attestation against the standards was carried out). In the statement of purposes, for example, there are references to professional learning and development and promoting quality (the development purpose) and also more accountability-oriented purposes such as determining if teachers should gain or retain a practicing certificate, strengthening confidence in the teaching profession and the reference to the essential nature of the knowledge and practices for effective teaching. The purposes described for the current standards are to

- describe the essential professional knowledge in practice and professional relationships and values required for effective teaching

Table 15.3 Standards for the teaching profession

The standards
• Te Tiriti o Waitangi partnership: Demonstrate commitment to tangata whenuatanga and Te Tiriti o Waitangi partnership in Aotearoa New Zealand
• Professional learning: Use inquiry, collaborative problem-solving and professional learning to improve professional capability to impact on the learning and achievement of all learners
• Professional relationships: Establish and maintain professional relationships and behaviors focused on the learning and wellbeing of each learner
• Learning-focused culture: Develop a culture that is focused on learning and is characterized by respect, inclusion, empathy, collaboration and safety
• Design for learning: Design learning based on curriculum and pedagogical knowledge, assessment information and an understanding of each learner's strengths, interests, needs, identities, languages and cultures
• Teaching: Teach and respond to learners in a knowledgeable and adaptive way to progress their learning at an appropriate depth and pace

Table 15.4 Elaborations of standards for the teaching profession

Elaborations
• Inquire into and reflect on the effectiveness of practice in an ongoing way, using evidence from a range of sources
• Critically examine how my own assumptions and beliefs, including cultural beliefs, impact on practice and the achievement of learners with different abilities and needs, backgrounds, genders, identities, languages and cultures
• Engage in professional learning and adaptively apply this learning in practice
• Be informed by research and innovations related to: content disciplines; pedagogy; teaching for diverse learners, including learners with disabilities and learning support needs; and wider education matters
• Seek and respond to feedback from learners, colleagues and other education professionals and engage in collaborative problem-solving and learning-focused collegial discussions

- promote high-quality teaching and leadership for all learners across all education settings
- set the standard expected for teachers to be issued with a practicing certificate
- provide a framework to guide our career-long professional learning and development as a teacher
- promote the status of the teaching profession through making explicit the complex nature of teachers' work
- strengthen public confidence in the teaching profession.

15.1.6.1 Increased Demand for Evidence

The *Standards for the Teaching Profession* specified that high-quality practices will generate naturally occurring evidence that can be used for discussion and analysis. It was made clear that for the purposes of appraisal, it was not expected that teachers would need to identify evidence of individual elaborations, but they were required to compile sufficient evidence of the quality of their practice to reflect the standard.

In practice, the standards together with the expectations for appraisal were understood by many educational leaders to mean that teachers were required to undertake intensive inquiries, professional development undertaken was to be thoroughly documented, and the portfolios of evidence compiled by teachers were to be extensive. In some schools, the demands on schools and teachers became excessive (Alison & Willetts, 2020). Though there was recognition by the Post-Primary Teachers' Association (PPTS) that while there were overly engineered processes occurring in some schools, that was not the case across the system:

> Teacher appraisal was identified by the 2016 Joint Working Group on Secondary Teacher Workload as a major driver of unnecessary work for many teachers….[various agencies] have been working to find ways of helping to reduce the unnecessary workload of teachers generated from over-engineered appraisal activities….Some schools currently do not have over-engineered appraisal systems and cause very little work for their teachers in this area. Other

> schools may have over-engineered their requirements on staff because of a misunderstanding about what was required of schools'

15.1.7 Abandoning Teacher Appraisal: The Turn to a Professional Growth Cycle (2019)

During teacher salary negotiations in 2019, an agreement was reached that appraisal for New Zealand teachers, and therefore the associated audits of appraisal, be abandoned. Teachers were also instructed they did no longer need to collect evidence about meeting the Standards for the Teaching Profession. An accord (PPTA et al., 2019), signed by the Secretary of Education and the New Zealand Post-Primary Teachers' Association (PPTA), the primary teachers' union and the New Zealand Education Institute: Te Riu Roa, (NZEI), set out that:

> Evidence shows that performance appraisal as an accountability instrument does not demonstrably lift teacher quality and contributes to a low trust high workload environment. As part of the accord implementation process the parties, NZSTA and the Teaching Council will work together to remove performance appraisal The Minister has committed to bringing forward legislation to remove the relevant requirements in legislation. (p. 2)

In some communications, the evidence part of the rationale for abandoning appraisal was framed in terms of an absence of evidence in support of appraisal. In their news to members, for example, the PPTA (2020) indicated "there is a lack of evidence that appraisal lifts teacher quality or improves student outcomes," whereas in the accord itself, it was framed in terms of the presence of negative impacts: "Evidence shows that performance appraisal as an accountability instrument does not demonstrably lift teacher quality and contributes to a low trust high workload environment." The former is likely a more defendable statement of the evidence situation. Other official statements made a stronger link between the particular type of appraisal and the absence of positive impacts:

> As part of the collective bargaining between the Government, PPTA and NZEI, an Accord was developed by the Secretary for Education, the NZEI, and the PPTA. As part of the Accord, the parties, along with New Zealand School Trustees Association and the Teaching Council, agreed to remove the requirement for appraisal of teachers. This was in recognition that compliance-driven appraisal used as an accountability instrument does not demonstrably lift teacher quality, and has instead contributed to a low trust, high workload environment.

These distinctions in the range of claims made during this process are important, since the warrant for the various claims used in relation to calls to remove teacher appraisal is quite different:

Claim 1: appraisal ***does not impact positively*** on teaching and learning.
Claim 2: there is ***little evidence*** that appraisal does ***impact positively*** on teaching and/or learning.
Claim 3: appraisal ***impacts negatively*** on teaching and learning.

> Claim 4: appraisal ***of a type that is compliance-driven and used (only) as an accountability instrument does not impact positively*** on teaching and learning.

Despite these subtly distinct claims, and varying robustness of the support for some of them, teacher appraisal was removed (unofficially in late 2019 in anticipation of the revised legislation, but took effect officially in 2020) and was treated as a victory from a teaching association/union point of view. Teacher appraisal, they claimed, had "contributed to a low trust environment, which is good for no one. Now is the time to move toward a high trust model" (PPTA, 2020). The removal of teacher appraisal occurred, therefore, in the highly politicized context of contract negotiations.

15.1.8 A Theory-of-Action Approach to Describing Policy Actions

To understand the shift away from appraisal, to a new process referred to as the professional growth cycle, I take a theory-of-action approach to describing both the former and current (2021) approaches. Attending to the theories of action for both allows consideration of the extent to which the policy shift involved single- or double-loop learning as explained below. First, the particular notion of a theory of action that I use for this analysis, based on the work of Argyris and Schon, is explained (Argyris & Schon, 1974, 1996). Theories of action have three components. The first is the observed actions; a second is the constraints that rule in those actions and rule others out—the actors' values and associated beliefs. The third component is the consequences of those actions including both intended and unintended. Theory-of-action approaches have been used in a range of research contexts including for research focused on understanding collaboration (Sinnema et al., 2021), in-service teacher education (Peeters & Robinson, 2015), reporting to parents (Hannah et al., 2018) and on-the-job decision-making (Robinson & Donald, 2014). Here, we use it in the policy context, foregrounding the appraisal-related actions of policy makers and the consequences of those actions for teachers, schools and the children and young people who they are responsible for.

15.1.8.1 A Theory of Action to Describe the Appraisal Approach up to 2019

In line with the theory-of-action approach introduced above, Fig. 15.1 details the approach to teacher appraisal introduced in 1996 that was in place until 2020 by foregrounding the actions of policy makers. While there are choices in explaining practice using a theory-of-action account, here I foreground policy/policy makers actions collectively. I treat the actions of practitioners, including educational leaders and teachers as central to the consequences in the theory of action; these include both intended and unintended consequences.

Governing variables	Actions	Consequences
1) ***Entitlement***: *Every teacher should each have an annual appraisal in relation to agreed - performance objectives (or goals),* 2) ***Improvement and accountability Formative and Summative purposes -:*** *Appraisal should deal with the improvement of teaching and learning in both the formative and summative sense* *i) [appraisal should] "provide a positive framework for improving the quality of teaching (and therefore learning) in New Zealand schools."* *ii) [appraisal should] fulfil both an accountability (summative) and development (formative) function:* *iii) Appraisal should enable appraisers to establish if the standards teachers are accountable for are being met* *iv) Minimum standards of accountability and quality assurance should be set out, that are flexible enough for boards of trustees to establish appraisal systems appropriate to their school and community* 3) ***Prescription:*** *Requirements should be set out for appraisal in relation to elements, foci and principles* 4) ***Integration:*** *Appraisal should be integrated into schools' planning and organisation for professional development.* 5) ***Goal-driven:*** *school-wide goals and objectives should be linked with the professional goals of individuals* 6) ***Autonomy:*** *Schools should have autonomy to design their own appraisal approaches* ***But*** 7) ***Variable capability:*** *There is variable capability for and understanding about high quality approaches* ***so*** 8) ***External Scrutiny:*** *Auditing processes will ensure the quality of appraisal* 9) ***Evidence.*** *Evidence is required to provide assurance of the quality of appraisal processes and decisions*	**Convey requirements for schools' approach to appraisal** 1. Require that teachers are appraised annually in relation to agreed performance objectives 2. Set out requirements for schools' appraisal processes in relation to: - **Elements to be included:** observation, self-appraisal, discussion with between appraiser/appraisee and with a written report - **Foci of appraisal :** in relation to **key performance areas** (1996) of teaching, school-wide and management responsibilities and from 1998 **against professional standards**; - **Consequences** of summative function: Attestation against professional standards established during appraisal linked to making decisions on salary progression. - **Principles of appraisal policies and procedures** should: - *Be part of an integrated performance management system operating within the school;* - *be appropriate to individual teachers, the school, and the wider community;* - *be developed in a consultative manner with teachers;* - *be open and transparent;* - *have a professional development orientation;* - *be timely and helpful to the individual teacher; and* - *give consideration to matters of confidentiality, including the provisions of the Privacy Act and the Official Information Act;* **Provide professional support** - Ruia - Appraisal project **Audit:** Compliance with the above should be scrutinized through external auditing of schools' appraisal Teaching council delegated to the Education Review Office to audited and required evidence as part of that process **Require evidence** Teaching council required teachers to compile evidence/portfolios	**The quality of approaches to appraisal were variable.** **Some/many schools/leaders instituted processes of teacher appraisal that were experienced by teachers as burdensome** - met requirements though exceeded accountability requirements and downplayed development purposes involved - extensive extended often individual inquiries - extensive documentation of evidence in paper and/or online portfolios and was - intensive, - over-engineered - time consuming - compliance oriented - Added weight and complexity. **Some/many educators felt distrusted** **Educators… (some/many) believed that appraisal did not impact positively on them or their learners** - unproductive - Reduced time for teaching - did not translate into professional development. - did not impact on student learning **Educators… (some/many) believed that appraisal impacted negatively on teachers** - linked to concerns about workload, wellbeing and other workforce issues **Commercial evidence portfolio systems proliferated** **Little robust evidence to give insights into the nature and extent of the impact of appraisal on the quality of teaching or students learning and progress.**

Fig. 15.1 1996–2020 Teacher appraisal policy theory of action

15.1.8.2 The Turn to a Professional Growth Cycle in Place of Teacher Appraisal (2020)

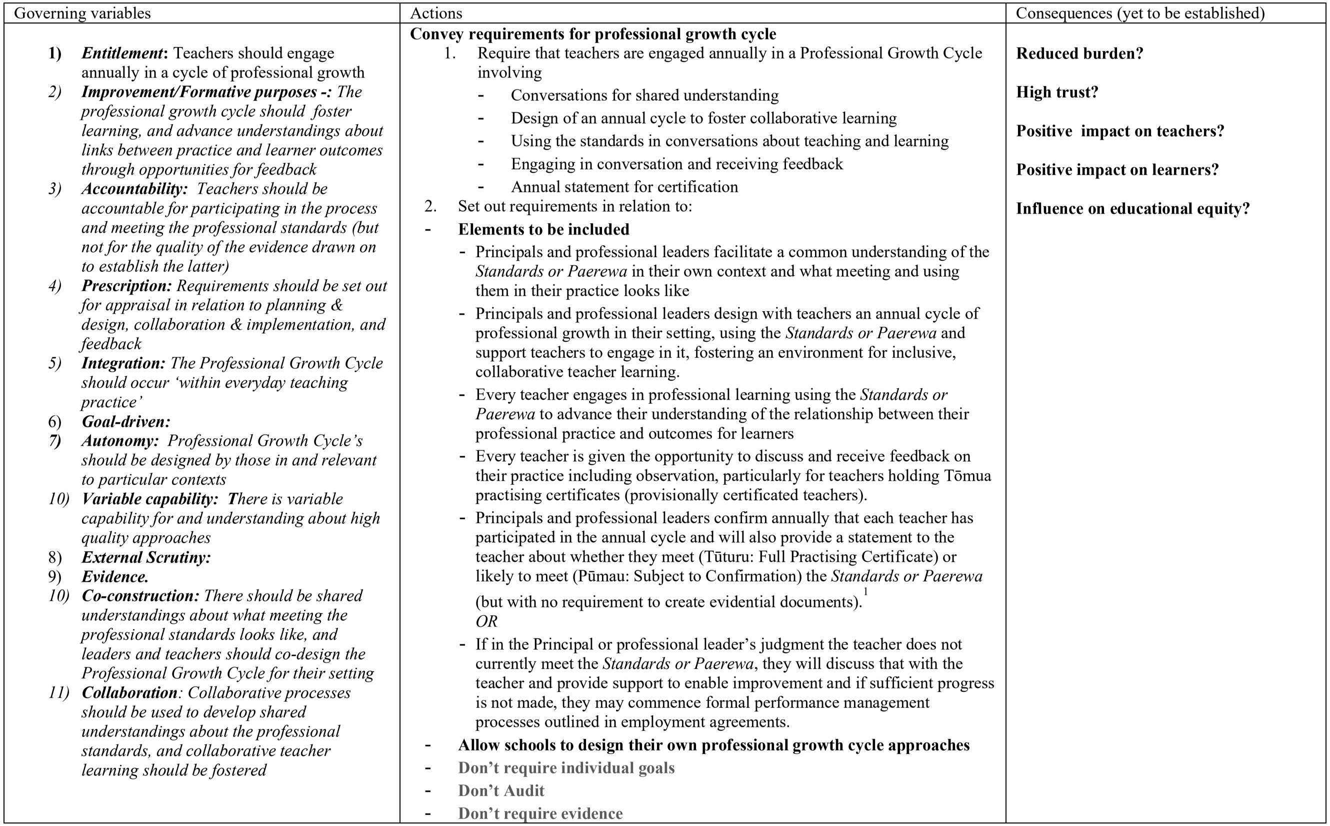
In 2019, it was announced that appraisal would be removed the following year and replaced by an alternative process named the professional growth cycle for teachers. Like appraisal, the professional growth cycle refers to the current professional standards, is required to occur annually and involves observation and feedback. Unlike the prior system of appraisal, the professional growth cycle is framed (as the term "growth" in its name suggests) in a way that foregrounds its formative and developmental purposes. The new process also emphasizes a more inclusive, collaborative approach and requires a process whereby teachers work together to understand the standards and what meeting them involves and to co-construct with their leaders the growth cycle for their context. Unlike the appraisal policy it replaces, the professional growth cycle does not require goal setting for individual teachers. The elements required in a professional growth cycle, and other actions required, together with an indication of constraints likely to drive those actions are set out in the theory of action on Fig. 15.2. Note that many of the constraints that underpinned actions of the previous (appraisal) system remain, some are no longer at play (indicated by the text with a strikethrough: goal-driven, external scrutiny and evidence), and others are new (collaboration and co-construction).

15.2 Critiquing the Removal of Teacher Appraisal

In this section, I critique the removal of teacher appraisal. This is not to suggest that the issues relating to workload, burden, trust and productivity did not deserve attention. Rather, I propose that appraisal could have been the target for improvement rather than removal and that such improvement would have been consistent with the notion of a system learns (Ministerial Advisory Group on Curriculum Progress and Achievement, 2019) that is key to policy initiatives in the New Zealand context.

15.2.1 Did Appraisal Deserve to Be Removed and Replaced?

I argue that the New Zealand system did not necessarily need to be replaced by a new process in order to bring teachers' professional growth into focus. That is not to say that features of the new professional growth cycle are not worthy. Rather, I suggest that there was an opportunity to examine the constraints driving problematic implementation of the previous appraisal process and to alter and refine both those constraints and related actions in ways that would address the weaknesses in the previous appraisal system. The suggestion for more attention to constraints is consistent with the notion of double- (rather than single-) loop learning.



Governing variables	Actions	Consequences (yet to be established)
1) ***Entitlement***: Teachers should engage annually in a cycle of professional growth *2)* ***Improvement/Formative purposes -:*** *The professional growth cycle should foster learning, and advance understandings about links between practice and learner outcomes through opportunities for feedback* *3)* ***Accountability:*** *Teachers should be accountable for participating in the process and meeting the professional standards (but not for the quality of the evidence drawn on to establish the latter)* *4)* ***Prescription:*** *Requirements should be set out for appraisal in relation to planning & design, collaboration & implementation, and feedback* *5)* ***Integration:*** *The Professional Growth Cycle should occur 'within everyday teaching practice'* 6) ***Goal-driven:*** *7)* ***Autonomy:*** *Professional Growth Cycle's should be designed by those in and relevant to particular contexts* *10)* ***Variable capability:*** ***T****here is variable capability for and understanding about high quality approaches* 8) ***External Scrutiny:*** 9) ***Evidence.*** *10)* ***Co-construction:*** *There should be shared understandings about what meeting the professional standards looks like, and leaders and teachers should co-design the Professional Growth Cycle for their setting* *11)* **Collaboration**: *Collaborative processes should be used to develop shared understandings about the professional standards, and collaborative teacher learning should be fostered*	**Convey requirements for professional growth cycle** 1. Require that teachers are engaged annually in a Professional Growth Cycle involving - Conversations for shared understanding - Design of an annual cycle to foster collaborative learning - Using the standards in conversations about teaching and learning - Engaging in conversation and receiving feedback - Annual statement for certification 2. Set out requirements in relation to: - **Elements to be included** - Principals and professional leaders facilitate a common understanding of the *Standards or Paerewa* in their own context and what meeting and using them in their practice looks like - Principals and professional leaders design with teachers an annual cycle of professional growth in their setting, using the *Standards or Paerewa* and support teachers to engage in it, fostering an environment for inclusive, collaborative teacher learning. - Every teacher engages in professional learning using the *Standards or Paerewa* to advance their understanding of the relationship between their professional practice and outcomes for learners - Every teacher is given the opportunity to discuss and receive feedback on their practice including observation, particularly for teachers holding Tōmua practising certificates (provisionally certificated teachers). - Principals and professional leaders confirm annually that each teacher has participated in the annual cycle and will also provide a statement to the teacher about whether they meet (Tūturu: Full Practising Certificate) or likely to meet (Pūmau: Subject to Confirmation) the *Standards or Paerewa* (but with no requirement to create evidential documents).[1] *OR* - If in the Principal or professional leader's judgment the teacher does not currently meet the *Standards or Paerewa*, they will discuss that with the teacher and provide support to enable improvement and if sufficient progress is not made, they may commence formal performance management processes outlined in employment agreements. - **Allow schools to design their own professional growth cycle approaches** - **Don't require individual goals** - **Don't Audit** - **Don't require evidence**	**Reduced burden?** **High trust?** **Positive impact on teachers?** **Positive impact on learners?** **Influence on educational equity?**

Fig. 15.2 2020 The professional growth cycle that replaced appraisal theory of action

15.2.2 Single- and Double-Loop Learning

Double-loop contrasts single-loop learning because it demands, for the purpose of improving organizational learning (and in this case, system learning) that individual and shared beliefs and assumptions that guide behavior are examined (Argyris & Schon, 1996). Double-loop learning, as Argyris explains, involves questioning the "fundamental design, goals, and activities of their organizations" (Argyris, 1976, p. 367). In single-loop learning, participants are encouraged to learn, but they must do so without questioning or seeking to change such fundamental elements of the organization. As Robinson explains "the capacity to double loop learn, and thus to question our assumptions about what counts as effective action, is essential if individuals and organizations are to detect and correct errors which are caused not simply by poor choice of strategy but by taken-for-granted values and assumptions" (Robinson, 2014, p. 755).

Attention to constraints that persisted, were removed and were added. In the section below, I highlight (a) the opportunity there was for new or revised constraints within a continued appraisal process, (b) that some of the problematic constraints (taken-for-granted assumptions that impacted on appraisal action) persist in the new professional growth cycle despite the process itself being new and (c) the risk that some of the constraints driving the previous appraisal process that have been abandoned might reduce the likelihood of success of the new professional growth process.

The potential for new or revised constraints instead of abandoning appraisal. Some of the constraints driving the actions required in the "new" professional growth cycle could potentially have been applied to appraisal. Take, for example, co-construction. The new initiative recognises the importance of leaders and teachers sharing understandings about what it means to meet the professional standards. It focuses on them co-constructing the process (professional growth cycle) for their setting. These are promising suggestions, and important given what evidence increasingly reveals about the value of such co-construction and collaboration. There could, however, have been moves to require the shared understandings, co-construction and collaboration within the context of the pre-existing appraisal process.

Furthermore, some would argue that notions of co-construction and collaborative teacher learning were at least permissible and arguably promoted given the emphasis on those in the professional standards in place at the time that were linked to teacher appraisal. In the new professional growth cycle, these are much more prominent, but I would argue teacher appraisal need not have been abandoned to make them so. The findings of the Education Review Office referred to earlier that there were pockets of high-quality practice with regard to appraisal (e.g., appraisal that was not burdensome, or unproductive, or overly engineered) suggest there was expertise in relation to appraisal that could have been leveraged with such collaboration, in order to spread good appraisal practice rather than remove the requirement for any appraisal practice given the weaknesses of some.

The persistence of constraints associated with problematic actions and consequences in the (old) appraisal system. Some of the (problematic) constraints that were associated with problematic actions and outcomes in the previous system persist, meaning that the new process may be susceptible to the same issues as appraisal was. For example, the assumption that there was sufficient capability for and understanding about high-quality approaches to appraisal (or at least access to opportunities to develop that capability) may explain how the actions resulted in implementation of variable quality. Capability and understanding are just as critical to the new professional growth cycle as it was to appraisal—therefore, changing the name and specifications in the process may not result in improvement if the constraint relating to capability is not addressed. What I am suggesting here is that giving greater attention, effort and resource to building capability for appraisal may have, at least in part, contributed to the improvement of appraisal in ways that impacted positively on the consequences of it. And similarly, not addressing levels of capability for these kinds of processes in schools has the potential to lead to professional growth cycle approaches that are equally as variable in quality as appraisal was. In other words, the change may not solve the problems it is intended to solve.

Similarly, the constraint of autonomy, the freedom and expectation for schools to design their own processes (a constraint that was in the appraisal theory of action and is retained in the professional growth cycle theory of action) could lead to professional growth cycles that are also burdensome and overly engineered just as some appraisal processes were (or were claimed to be).

Abandoned constraints that risk reducing the likelihood of success. An implicit belief guiding a key element of the discontinued appraisal process was that goal setting had a role to play in the process. The requirement for teachers to set goals to be the focus of their appraisal efforts recognized the theoretical rationale for goal setting as related to motivation, productivity and performance (Locke & Latham, 2012) and empirical work in educational settings showing that increased effort and commitment (to goals set in appraisal processes) are associated with increased goal achievement (Sinnema & Robinson, 2012).

15.2.3 Assumptions Worth Testing in the Rationale for the Shift

The shift away from teacher appraisal to a new professional growth cycle implies a number of assumptions that ought to be tested as the change embeds and develops in the New Zealand system.

One assumption is that the appropriate solution to the problem of appraisal being burdensome was removing rather than improving appraisal itself. As discussed earlier, that assumption can be challenged. The presence of high-quality approaches in some schools, with some schools demonstrating appraisal processes that were not burdensome and did build trust (as highlighted by the review of appraisal across many school), suggests that such quality approaches were, under particular conditions, feasible. It follows that learning about those conditions and leveraging the expertise involved would have allowed the system to learn rather than switch.

Another assumption was that addressing the negative impacts on teachers (burden, workload, etc.) attributed to the auditing and monitoring of appraisal required removing rather than a) improving the approach to monitoring/auditing and b) educators' understandings of the expectations under which they were being audited. The principle of co-construction evident in the new professional growth cycle approach could have been used, for example, to (re)co-construct the indicators used to audit and emphasize an improvement orientation of audits. This could potentially have steered more attention to the quality of appraisal including quality in terms of the appropriate levels of burden/productivity/reward experienced by teachers in their appraisals. It is a shame for the system to have lost the role of audits in creating a shared understanding about areas for improvement in appraisal processes (including "sharpening the focus on students at risk of not achieving; providing more purposeful observations of practice and more useful feedback/feed forward; deepening the quality of goal setting and self-reflection; refining and strengthening 'teaching as inquiry'; and considering how to strengthen teachers' cultural competencies" (The New Zealand Education Council, 2018, p. 5). A third implicit assumption in statements about the rationale for shifting from appraisal to a professional growth cycle is that robust appraisal processes and high degrees of trust cannot co-exist. That assumption is not necessarily true. It could be possible, for example, for the approach to appraisal to integrate the task of ensuring and promoting improvement in teaching and learning, while developing and sustaining quality relationships and a climate of trust. To do so would require conditions conducive to such integration including opportunities for appraisers to develop the capabilities required to attend to both. In addition, it is apparent that while the new process of professional growth is not explicitly referred to as serving accountability purposes, the part of the cycle that requires principals to confirm that each teacher not only participated in the cycle but also met the professional standards suggests an accountability function though one that is not explicitly identified as such. This could, potentially, lead to the opposite of what is intended and decrease rather than strengthen trust in the process.

Statements about the rationale for the shift away from appraisal also suggested that there was no evidence for the impact of appraisal on student learning. A misguided assumption related to that is that an absence of evidence of the relationship means that the relationship between appraisal and student learning does not exist. A lack of such evidence does not so much support the decision to remove appraisal, but strengthens the case for high-quality research to determine and understand the relationship.

In summary, while the new system is still recent and the consequences of it are uncertain, I argue that much of what the new system emphasizes could have been applied to an improved appraisal process. The aspirations for removing unproductive compliance activity, reducing workload and focusing on professional growth were possible within an appraisal framework. There was also potential, I suggest, to retain an accountability mechanism in appraisal alongside moves to increase trust and reduce burden. This would have involved altering the approach to and focus of accountability; holding schools accountable, for example, for not implementing unproductive levels of compliance activity, for designing appraisal approaches that are not burdensome, that increase rather than reduce teacher trust, and at the same time support the improvement of teaching and learning. Such integration of accountability alongside improvement purposes might have been positioned as professionalizing rather than de-professionalizing. Furthermore, it is particularly important in systems seeking to address issues of educational inequity, such as New Zealand (The New Zealand Ministry of Education, n.d.-a). If the decision to remove appraisal from this system is considered from a political point of view, with the success of contract negotiations at the center of decision-making, the move may be considered a success. However, from the perspective of students who are most disadvantaged by educational inequity and who have most to gain from system mechanisms that ensure improvements in teaching, the removal of appraisal in the New Zealand education system is, in my view, a significant loss.

References

Aitken, G., & Sinnema, C. (2008). *Effective pedagogy in social sciences/tikanga ā iwi: Best evidence synthesis iteration.* Ministry of Education.

Alison, J., & Willetts, R. (2020, July 14). Opportunity missed: Why the government's failure to reform tomorrow's schools means some schools will continue to make poor decisions, with negative impacts on teachers and students. *Ipu Kererū: Blog of the New Zealand Association for Research in Education.* https://nzareblog.wordpress.com/2020/07/14/tomorrows-schools/

Argyris, C. (1976). Single-loop and double-loop models in research on decision making. *Administrative Science Quarterly*, *21*(3), 363–375. JSTOR. https://doi.org/10.2307/2391848

Argyris, C., & Schon, D. (1974). *Theory in practice: Increasing professional effectiveness.* Jossey-Bass.

Argyris, C., & Schon, D. (1996). *Organizational learning II: Theory, method and practice.*

Education Review Office. (2012). *Teaching as inquiry: Responding to learners.* Crown.

Education Council of Aotearoa New Zealand. (2017). Our code, our standards: Code of professional responsibility and standards for the teaching profession. Available at: https://teachingcouncil.nz/assets/Files/Code-and-Standards/Our-Code-Our-Standards-Nga-Tikanga-Matatika-Nga-Paerewa.pdf

Government of New Zealand. (1989a). *Education act*. Government Printer.

Government of New Zealand. (1989b). *State sector amendment Ac*. Government Printer.

Government of New Zealand. (1989c). *Tomorrow's schools: The reform of education administration in New Zealand*. Government Printer.

Hannah, D., Sinnema, C., & Robinson, V. (2018). Theory of action accounts of problem-solving: How a Japanese school communicates student incidents to parents. *Management in Education, 33*(2), 62–69. https://doi.org/10.1177/0892020618783809

Lai, M., & Sinnema, C. (2022). Evidence use in education in Aotearoa New Zealand. In J. Malin & C. Brown (Eds.), *The Emerald international handbook of evidence-informed practice in education: Learning from international contexts.* Emerald Publishing Ltd.

Locke, E. A., & Latham, G. P. (2012). *New developments in goal setting and task performance.* Taylor & Francis Group. http://ebookcentral.proquest.com/lib/auckland/detail.action?docID=1104793

Ministerial Advisory Group on Curriculum Progress and Achievement. (2019). *Strengthening curriculum, progress, and achievement in a system that learns: Report by the Curriculum, Progress, and Achievement Ministerial Advisory Group to the Minister of Education.*

Peeters, A., & Robinson, V. (2015). A teacher educator learns how to learn from mistakes: Single and double-loop learning for facilitators of in-service teacher education. *Studying Teacher Education, 11*(3), 213–227. https://doi.org/10.1080/17425964.2015.1070728

Piggot-Irvine, E. (2000). Appraisal—The impact of increased control on the "state of play" in New Zealand schools. *Journal of Educational Administration, 38*(4), 331–351. https://doi.org/10.1108/09578230010373606

PPTA. (2020, July). *The moratorium on appraisal*. https://www.ppta.org.nz/news-and-media/moratorium-on-appraisal-july-2020/

PPTA, NZEI, & Secretary for Education. (2019). *Accord between the Ministry of Education, NZEI Te Riu Roa and PPTA Te Wehengarua*. https://www.initials.org.nz/collective-agreements/document/850

Robinson, V. (2014). Single and double loop learning. In *Encyclopaedia of educational theory and philosophy*. Sage.

Robinson, V., & Donald, R. (2014). On the job decision-making: Understanding and evaluating how leaders solve problems. In *Decision-making in educational leadership. Principles, policies, and practices* (pp. 93–108). Taylor & Francis.

Sinnema, C. (2005). *Teacher appraisal: Missed opportunities for learning* [Unpublished doctoral thesis]. The University of Auckland.

Sinnema, C. (2015). The ebb and flow of curricular autonomy: Balance between local freedom and national prescription in curricula. In D. Wyse, L. Hayward, & J. Pandya (Eds.), *The Sage handbook of curriculum, pedagogy and assessment* (Vol. 2, pp. 965–983). Sage.

Sinnema, C., & Aitken, G. (2011). Teaching as inquiry in the New Zealand curriculum: Origins and implementation. In J. Parr, H. Hedges, & S. May (Eds.), *Changing trajectories of teaching and learning* (pp. 29–48). NZCER.

Sinnema, C., & Aitken, G. (2013). Emerging international trends in curriculum. In M. Priestley & G. J. J. Biesta (Eds.), *Reinventing the curriculum: New trends in curriculum policy and practice*. Bloomsbury Academic.

Sinnema, C., & Aitken, G. (2019). Teaching as Inquiry. In M. Hill & M. Thrupp (Eds.), *The professional practice of teaching in New Zealand* (6th ed.). Cengage.

Sinnema, C., Hannah, D., Finnerty, A., & Daly, A. J. (2021a). A theory of action account of an across-school collaboration policy in practice. *Journal of Educational Change, 33*(2), 62–69.

Sinnema, C., Liou, Y.-H., Daly, A., Cann, R., & Rodway, J. (2021b). When seekers reap rewards and providers pay a price: The role of relationships and discussion in improving practice in a community of learning. *Teaching and Teacher Education, 107*, 103474. https://doi.org/10.1016/j.tate.2021.103474

Sinnema, C., & Robinson, V. (2012). Goal setting in principal evaluation: Goal quality and predictors of achievement. *Leadership and Policy in Schools, 11*(2), 135–167.

The New Zealand Education Council. (2018). *Appraisal as a catalyst for improved learner outcomes*. https://ero.govt.nz/our-research/appraisal-as-a-catalyst-for-improved-learner-outcomes-two-years-on

The New Zealand Education Review Office. (2014). *Supporting school improvement through effective appraisal*. https://thehub.swa.govt.nz/resources/supporting-school-improvement-through-effective-teacher-appraisal/

The New Zealand Ministry of Education. (n.d.-a). *PISA 2018 New Zealand summary report: System performance and equity*. The New Zealand Ministry of Education. https://www.educationcounts.govt.nz/__data/assets/pdf_file/0006/196629/PISA-2018-NZ-Summary-Report.pdf

The New Zealand Ministry of Education. (n.d.-b). *Ruia: Teacher appraisal for Maori learners' success*. https://appraisal.ruia.educationalleaders.govt.nz/

The New Zealand Ministry of Education. (1997). Performance management systems: PMS1: Performance appraisal. *Education Gazette*, *10th February Supplement*.

The New Zealand Ministry of Education. (1998a). Interim professional standards: Primary school deputy/assistant principals, primary school teachers. *The Education Gazette*, *10 February Supplement*.

The New Zealand Ministry of Education. (1998b). *Teacher performance management: Primary school teachers, primary school deputy/assistant principals: A resource for boards of trustees, principals, and teachers.*