# Validity Evidence of Internal Structure and Subscores Use of the Portfolio in the Chilean Teachers' Evaluation System

Dissertation

Presented in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy in the Graduate School of The Ohio State University

By

Pamela Soto Ramirez, M.S.

Graduate Program in Educational Studies

The Ohio State University

2020

Dissertation Committee

Dr. Jerome D'Agostino, Ph.D., Advisor

Dr. Jessica Logan

Dr. Shayne Piasta

Copyrighted by

Pamela Soto Ramirez

2020

Abstract

There is consensus around the fact that quality of teaching is one of the most important school-level factors to influence student achievement at school. Evidence from research suggests that better-qualified teachers can be the determining factor for student achievement and development (Jordan et al., 1997; Sanders & Rivers, 1996; Wright et al., 1997). Therefore, policy makers advocate for ongoing improvements in teacher quality variables, in which the implementation of a well-designed teacher evaluation system has been found to be one of the most effective ways to improve teacher quality (Darling-Hammond, 2010; Looney, 2011; Rockoff & Speroni, 2011). The case of Chile is a particularly interesting example of a teacher evaluation system since its implementation, a validation process that has included not only the experience and documentation regarding the process, but also a comprehensive agenda regarding the validity and reliability of the instrument and evaluation consequences (Taut & Sun, 2014).

The purpose of this dissertation is to contribute to the body of research on the evidence of the validity of Chilean teacher evaluation. Specifically, I focus on one of the teacher evaluation instruments: the portfolio. Through the portfolio, teachers provide evidence of their best practices in three modules: a set of pedagogical materials, video recording class, and collaborative work (not mandatory).

In order to accomplish this goal, I use the data from the portfolio results of the 2017 Chilean National Teacher Evaluation ($N$ = 21,982). I use descriptive statistics,

exploratory confirmatory factor analysis (ECFA), and factorial invariance to assess the structure of the portfolio across different teacher subgroups: teachers with and without the collaborative work module score, rural/urban teachers, and six different teaching levels. I also compare the theoretical weight assigned to each one of the portfolio indicators with the empirical data. Finally, I evaluate if the portfolio subscores for the different modules have added value over the total portfolio score reported. For the analysis, Stata v.14 and M*plus* 8 are used.

Four main findings for portfolio validity were found. First, results from the ECFA indicate the portfolio's theoretical structure fits the data well for the groups of teachers without collaborative work results. Conversely, for the teachers with the results from the collaborative work module, the data did not clearly support the theoretical dimensions proposed by the portfolio. The second finding was related to the portfolio structure invariance across teachers, depending on their school location and teaching level. The results showed a strong factorial invariance (invariant thresholds) across rural and urban teachers, and weak factorial invariance (factor loadings) across the six groups of teaching levels. The third finding showed differences between the theoretical weight assigned to each portfolio indicator and the empirical weight. Finally, the fourth finding showed that at least two of the three modules of the portfolio had added value over the total score.

These findings provide evidence of portfolio validity, indicating that the instrument is invariant for teachers working in different contexts. However, a revision of the collaborative work module and the theoretical weights is suggested.

## Dedication

To my two daughters, Magda and Emilia, and to my husband, Rodrigo. Thank you for being with me always.

Acknowledgments

I want to thank all of those who supported me during the last six years of my life. I want to thank my advisor, Dr. Jerry D'Agostino for his support and patience during my master's and Ph.D. Thanks for supporting me despite the distance and trusting in me. I also want to thank my dissertation committee members Dr. Shayne Piasta and Dr. Jessica Logan for their feedback throughout my dissertation writing process. Thanks Shayne for encouraging me to write and for providing me with opportunities to improve as a researcher.

My Ph.D. progress and completion was possible because I had the support of a great group of QREM colleagues and friends. Thanks to Mine, Susie, Robert, Menglin, Hui, Meng Ting, Yixi, James, Becky, and Christa. Thanks to my friend Johana for being a good friend and supporting me always.

Thanks to my friends in Columbus, Ohio. To my next-door neighbor Ana and Orlando, for their unconditional support. Thanks to Josh and Jessie for being so kind during my visits to Columbus.

Thanks to my friends in Chile that always encouraged and supported me. Thanks to Emma, for supporting me in this final process of dissertation writing, and for always being positive about my progress. To my colleagues and friends from the UC Survey Center, for being with me in this final period.

Thanks to my family. My mother for always being with me when I needed it. My sister and her family, for their support, no matter the distance. My mother in law, for joining us and making us feel at home. My husband Rodrigo for this wonderful and crazy idea of studying abroad. Thanks to my daughters, Magda and Emilia, for their infinite patience. Everything is better because I have you by my side.

This work was supported by the Chilean National Commission of Science and Technology (CONICYT), Becas Chile for Ph.D. abroad, grant # 72170119.

## Vita

| | |
|---|---|
| 2016……. | M.S. Educational Studies, The Ohio State University |
| 2008 to 2014 …… | Associate Researcher, Faculty of Economy, University of Chile |
| 2005 to 2008 …… | Professional, Measurement System of Educational Quality, Educational Ministry, Chile |
| 2003…… | Psychology, Pontifical University Catholic of Chile |

## Publications

Piasta, S.B., Farley, K.S., Mauck, S.A., **Soto Ramirez, P**., O'Connell, A.A., Schachter, R.E., Justice, L.M., Spear, C.F., & Weber-Mayrer, M. (2019). At-scale, state-sponsored language and literacy professional development: Impacts on early childhood practices and children's outcomes. *Journal of Educational Psychology*.

Piasta, S.B., **Soto Ramirez, P**., Farley, K.S., Justice, L.M., & Park, S. (in press). Exploring the nature of association between educators' knowledge and their emergent literacy classroom practices. *Reading and Writing*.

## Fields of Study

Major Field: Educational Studies

Table of Contents

List of Tables

List of Figures

# Chapter 1. Introduction

Several studies have confirmed that the quality of teaching is one of the most important school-level factors that influences student achievement at school (Chetty et al., 2011; Darling-Hammond, 2000; Gerritsen et al., 2017; Hattie, 2012; Rockoff, 2004; Sanders & Rivers, 1996). The evidence suggests that teachers make a difference, as better-qualified teachers can be the determining factor for student achievement (Jordan et al., 1997; Sanders & Rivers, 1996; Wright et al., 1997) and for the child's future development. For instance, a study carried out in the Chicago Public Schools indicates, one standard deviation in the quality of Math teachers' improvement has been associated with an overall increase of student math scores by approximately one-fifth of average yearly gains over the course of one year (Aaronson et al., 2007). Furthermore, research that used the data from the University of Texas at Dallas (UTD) Texas Schools Project indicated that the increase in one standard deviation of the teacher quality distribution has a larger effect on reading and math students' achievement than the reduction in class size by ten students (Rivkin, Hanushek, & Kain, 2005). Research has also shown that students from low-income families may benefit from very effective teachers and may potentially achieve the same levels as their peers from high-income families (Looney, 2011).

Although all of the studies indicate that good teaching is important, it is less clear what variables are involved in defining a quality teacher (Goldhaber, 2002). Teacher quality is a complex phenomenon, and there is little consensus on the definition and the

way that it can be measured (Heck, 2009). Different studies have found that one of the most effective ways to improve teacher quality is to implement a well-designed teacher evaluation system that works in tandem with professional learning and development (Darling-Hammond, 2010; Looney, 2011; Rockoff & Speroni, 2011). Teacher assessment makes it possible to describe their performance, identifying main strengths and weaknesses, and setting out their training and needs support (Schmelkes, 2015). Therefore, timely and informative feedback is vital to any effort in improving teacher quality, and evaluation systems serve the purpose of providing feedback and guidance to teachers to improve their professional practices (Tucker & Stronge, 2005). The relevance of teacher evaluation systems is reflected in the increasing number of countries that include teacher evaluation in their National Education Agendas (Organization for Economic Co-operation and Development [OECD], 2013). In fact, out of the 28 countries surveyed by the Organization for Economic Co-operation and Development (OECD) in the Review on Evaluation and Assessment Frameworks for Improving Outcomes, 22 reported having national-level policy frameworks for teacher evaluation, and the six remaining countries reported well-designed and implemented practices to provide feedback on teachers' work (Huber & Skedsmo, 2016).

Given the importance of teacher evaluation systems, assessments used to evaluate teachers should yield scores that reflect the underlying construct of teaching quality (Looney, 2011). Good evaluation systems use instruments that have technical validity and protect the integrity of the evaluation process (Bruns & Luque, 2014). Teaching and learning will not improve if we fail to give teachers high-quality feedback based on accurate assessments of their instruction (T. J. Kane et al., 2014). The *Standards for*

*Educational and Psychological Testing* (American Educational Research Association [AERA], American Psychological Association [APA], and National Council on Measurement in Education [NCME], 2014) suggest that we have a professional responsibility to engage with and monitor the validity evidence for any large-scale testing and examination system, calling for a comprehensive validation agenda, especially for those systems with high-stakes consequences. It is imperative that teacher evaluation systems receive sufficient investment in the design and evaluation to ensure that they capture the main elements of teacher quality. Without the most effective evaluation tools and processes, teacher evaluation will have little impact on these systems (Looney, 2011).

Nevertheless, comprehensive validation of assessment systems is not an easy task due to the fact that the systems are complex and diverse (Taut et al., 2012). Thus, this may explain why there is little comparative information on the processes used to validate different teacher evaluation systems in different countries and the lack of well-documented literature related to comprehensive validation efforts for large scale assessment systems (Looney, 2011; Taut et al., 2012).

The case of Chile is a particularly interesting example of a teacher evaluation system since its implementation, a validation process carried out by researchers from the Measurement Center of the Pontifical Catholic University of Chile (MIDE UC). Under the supervision of the Ministry of Education of Chile, MIDE UC has been the institution in charge of implementing the Teacher Evaluation System since 2003 to the present. The validation process developed in MIDE UC included not only the experience and documentation regarding the process but also a comprehensive agenda regarding the

validity and reliability of the instrument and evaluation consequences (Taut & Sun, 2014).

Although the validation process was initially very successful in its implementation since 2012, there has been no systematic and organized validation agenda. Given this context, the MIDE UC Research team has recently called for a study group to promote research related to validating the teacher evaluation system in Chile, that includes from 2016 on a new Teacher Professional Development System that uses some of the same evaluation tools that have already been used by the Chilean Teacher Evaluation. Specifically, the goal is to focus on the validation of the evaluation process and the potential consequences for teacher professional development. The study group is made up of researchers from different disciplines and universities that meet with the purpose of designing, implementing, and publishing studies related to the recent policies of teacher professional development in Chile (Centro de Medición, MIDE UC, 2019).

The purpose of the present dissertation is to contribute to the body of research on the evidence of the validity of evaluation tools used in the Chilean Teacher Evaluation and in the Chilean Teacher Professional Development System. Specifically, I will focus on one of the evaluation tools that has been used in both evaluation systems: the portfolio. The portfolio is considered the core of the evaluation system, as it is the most complex part of the evidence that the teachers have to submit, contributes the most to the calculation of the overall score for the evaluation, and as previously mentioned, it is the only overlapping evaluation tool between the two systems that are currently in use in Chile for teachers evaluation.

## Background of the Study

**Chilean School Education System.** The school system in Chile is organized into three levels: pre-school education (children up to 5 years old), primary education (1st to 8th grade), and secondary (9th to 12th grade). Secondary schools in Chile offer three possible pathways for students starting in 11th grade: Humanities/Science (HS), Technical/Professional (TP), and the Arts (smallest in terms of enrollment). All of the different pathways aim to prepare students for higher education; however, TP corresponds to vocational education and training that also aims to prepare students for labor market entry, offering up to 35 different specializations in 15 occupational areas (Santiago & OECD, 2013).

Adult education is also offered in Chilean schools. This educational model is offered to youth and adults who want to start or complete their studies based on the National Curriculum of Chile. The purpose of this education format is to guarantee the Constitutional right to complete primary and secondary education for all (Santiago & OECD, 2013).

The National Curriculum of Chile was developed under the principles of the General Law of Education (2009), which outlines key knowledge, skills, attitudes, and learning objectives that students should develop in order to accomplish these goals. The key goals are intended to enable people to lead their lives fully, to actively participate within their communities, and to contribute to the development of the country (OECD, 2017).

Chile's education system has a segmented structure based on the market-oriented reforms of the 1980s. These reforms entailed the decentralization of public-school

management responsibilities to local governments (municipalities) and the introduction of a nationwide voucher program. As a result, the system combines public, private, and charter providers for all ages. Therefore, in Chile, there are four types of school providers: *Municipal schools*, public schools administered by the respective local governments or municipalities; *Private subsidized schools or charters schools*, schools administered by a private non-profit or for-profit organization that receives a public subsidy per student for the same amount as municipal schools; *Private non-subsidized schools*, schools administered by private non-profit or for-profit organizations that do not receive public subsidies; *Schools with delegated administration*, schools owned by the Ministry of Education and mostly offering technical-professional education whose administration is delegated to a public or private non-profit organization (Santiago & OECD, 2013). In 2017, 11,749 schools were registered in Chile, of which 44% were municipal schools, 50% charter schools, 5% private non-subsidized schools, and 1% schools with delegated administrations (Ministerio de Educacion et al., 2018).

Even though municipal and private subsidized schools receive public funding, they operate under different conditions. While private subsidized schools had the freedom until recently to select their students, municipal schools are required to admit all children. Private subsidized schools are allowed to charge tuition up to a certain amount, whereas municipal schools are only allowed to do so at the secondary level. Thus, attendance at different schools depends greatly on family income: students from the most disadvantaged families attend municipals schools in the largest numbers, private subsidized schools receive students from a wider range of backgrounds, and private non-

subsidized schools are mainly attended by students from high-income families (Santiago & OECD, 2013).

With respect to location, 70% of the schools in Chile are located in urban areas and 30% in rural areas. The rural schools are mostly small municipal schools managed by one or two teachers. The smallest rural schools are mostly concentrated in the poorest areas with fewer possibilities. Therefore, teachers in those schools are constantly faced with the need to adapt their classes to the rural educational environment and the particular challenges they face. Teachers in rural schools have to adapt their pedagogical practices to the local reality and to the opportunities that education in small schools present (Villarroel, 2003).

**The Teaching Profession.** In 2017, there were 235,527 teachers working in the Chilean school system. The distribution by type of school provider was: 44% working in municipal schools, 45% in charter schools, 10% in private non-subsidized schools, and 1% in schools with delegated administration. With respect to gender, 73% of the teachers were women in 2017. Nevertheless, this proportion of women varied according to the level (99% worked in pre-school education, 76% in primary education, and 55% in secondary) and to the job they do at school (74% classroom teachers, 76% heads of technical pedagogical units, 62% management positions, and 63% school principals). In 2017, 88% of teachers were working in urban schools and 12% in rural schools. (Ministerio de Educacion et al., 2018).

Most teachers (94%) by 2017, worked in a single school, and 39% worked 44 hours or more per week. The age distribution of the teachers varies across school providers, with the presence of a higher proportion of older teachers in the municipal

schools (21% of the teachers are older than 55 years old, compared to 13% in charter schools). The vast majority of teachers had a professional degree in education (around 95% for each type of school; Ministerio de Educacion et al., 2018), which is consistent with the fact that initial teacher education is a requirement to enter the teaching profession (Santiago & OECD, 2013).

Local governments (municipalities) and private providers are the institutions that employ teachers. The local government pays teachers a baseline salary, as well as a set of additional bonuses based on a series of additional criteria. In the private sector, the employers must guarantee this baseline salary, but they can choose to pay higher salaries as they see fit. The salary bonuses that benefit teachers are available to both municipal and charter school teachers. They can include experience, training, the difficulty of work conditions, performance, responsibility, and so on (Santiago & OECD, 2013).

## Teacher Evaluation System in Chile

Since the late 1990s, the educational policy in Chile has begun to increase the accountability of teachers' performance in public schools. From the mid-90s, evidence of low student learning results caused the authorities to hold teachers as partly responsible, putting pressure on the implementation of a teacher evaluation system. Initially, teachers rejected the implementation of a system, which gave rise to a long period of discussion and negotiations on whether and what kind of teacher evaluation system should be implemented in the country (Avalos & Assael, 2006).

In 1998 a commission was formed up of three main parties in the process: the teachers as represented by their Teacher Union, the Chilean Association of Municipalities as their employers, and the Ministry of Education that pays their salaries.

The three groups started to negotiate the design and implementation of a system that evaluates teachers' performance in public schools. The negotiations involved designing and approving an evaluation system that hinged on the purpose of the teacher evaluation, what kind of assessment criteria should be used, and what evaluation procedures would be considered appropriate (Avalos & Assael, 2006).

Therefore, the tripartite commission worked in designing and approving an evaluation system for teachers that took into consideration a formative purpose, standards appropriate to school level and teaching experience, and the use of a wide array of evidence gathering procedures in order to discern which of these would work best. The results of the complex negotiation and consultation process, involving the three aforementioned parties, all of whom had to make concessions for a final agreement to be reached was the Chilean National Teacher Evaluation System (Tornero & Taut, 2010).

The system established in 2003 constituted a completely new development for the country's teaching force, aligning itself with current international thinking on teacher evaluation. It was considered as a comprehensive, compulsory teacher evaluation system for all of the teachers who worked in municipal schools. The initial goal of the teacher evaluation was to improve teachers' practices and to promote their continuous professional development, always keeping student learning central to the mission (Bonifaz, 2011).

An important element during the negotiation process was around the need to have assessment criteria that aligned itself with a national framework that defined standards for the teaching profession to provide clarity on expectations for the profession. The Ministry of Education took the lead in this, producing a set of criteria based on the work of

Danielson (2007), the Praxis III (Educational Testing Service [ETS]), and the standards for initial teacher education elaborated by the Ministry of Education (Ministerio de Educacion, 2003; Assaél & Pavez, 2008). This framework was discussed, analyzed, and approved in two successive public meetings with the teachers. The discussion at the teacher meetings resulted in the creation of these standards, culminating in the final document: the Good Teaching Framework (GTF; *Marco para la Buena Enseñanza, in Spanish*; Ministerio de Educacion de Chile, 2008).

This framework provides a clear expectation of what teachers should know and be able to do as part of their daily work in and outside of the classroom. Following Danielson's framework that identifies those aspects of a teacher's responsibilities that have been documented as promoting improved learning through empirical and theoretical research (Danielson, 2007), GTF specifies domains, criteria within domains, descriptors for each criterion, and performance levels for descriptors that outline teacher responsibilities. The GTF contains 20 criteria grouped into four domains specific to the task of teaching: preparing for teaching, creating a learning environment, opportunity to learn for all students, and professional responsibilities. Each criterion is classified by performance levels: unsatisfactory, basic, competent, and outstanding, which are written in behavioral language that allows teachers to translate the standards into actual events in the classroom or in instructional planning (OECD, 2013). The same GTF framework applies to all teachers who are taking part in the evaluation without differentiating between different school levels or subject areas. Table 2.1 provides the list of domains and criteria for the GTF.

Table 1.1.
*Domains and Criteria of the Good Teaching Framework (GTF)*

| Domains | Criteria (the teacher should be prepared to:) |
|---|---|
| Domain A: Preparing for Teaching | A1. Masters the subjects taught and the national curricular framework.<br>A2. Knows the characteristics, knowledge, and experiences of his/her students.<br>A3. Masters the pedagogy of the subjects or disciplines they teach.<br>A4. Organizes the objectives and contents consistent with the curricular framework, and the characteristics of particular students.<br>A5. Uses assessment strategies that are consistent with the learning objectives, the subject taught, and the national curricular framework, and allows all students to show what they have learned. |
| Domain B: Creating a Learning Environment | B1. Creates an environment of acceptance, equality, trust, solidarity, and respect.<br>B2. Shows high expectations for students' ability to learn and develop themselves.<br>B3. Creates and implements consistent classroom rules.<br>B4. Creates an organized working atmosphere that provides the spaces and resources necessary for learning. |
| Domain C: Opportunity to Learn for All Students | C1. Communicates the learning objectives in a clear and accurate way.<br>C2. Plans challenging and consistent teaching strategies that are relevant for the students.<br>C3. Focuses on the most important content and uses terms that students are able to understand.<br>C4. Optimizes teaching time.<br>C5. Promotes critical thinking.<br>C6. Evaluates and monitors the learning process and the absorption of the material by the students. |

| | |
|---|---|
| Domain D: Professional Responsibilities | D1. Regularly reflects on his/her teaching skills.<br>D2. Builds a professional and collaborative relationship with his/her peers.<br>D3. Takes on advising responsibilities.<br>D4. Promotes respect and works collaboratively with the students' parents or guardians.<br>D5. Updates information relevant to the teaching profession, the educational system, and the current policies. |

*Note.* Translated and adapted from Ministry of Education (2008), Marco para la Buena Enseñanza (Good Teaching Framework), CPEIP, Santiago, ww.docentemas.cl/docs/MBE2008.pdf.

Through this evaluation system based on the teachers' abilities to achieve GTF standards (OECD, 2017), teacher performance is assessed every four years, according to four evaluation tools:

1) *Self-evaluation*, a structured questionnaire organized according to the four domains of the GTF. Its objective is to generate teachers' reflection on their own practices. The teacher rates his or her performance in 12 proposed areas. Teachers also have the possibility of adding information about the context of their teaching.

2) *Peer evaluation*, a trained classroom teacher, from the same discipline and grade level, assesses the teacher using a standardized set of six questions that covers the domains from the GTF.

3) *Third-party assessment report,* a structured questionnaire that is completed by both the school principal and the dean of academics of the school, covering a range of areas of the teacher's professional activity according to the GTF.

4) *Teacher performance portfolio*, designed for teachers to provide evidence of their best pedagogical practices. The portfolio is prepared for a particular grade level and subject area; however, it is standardized to make the evaluation experience comparable across subjects and grade levels. Teachers are provided with a manual, which describes the dimensions and indicators evaluated in the portfolio (Flotts & Abarzua, 2011).

The performance level rating for each indicator in each one of the four evaluation tools previously described is based on specifically designed assessment rubrics that evaluate the teacher within four possible levels: unsatisfactory, basic, competent, and outstanding. These rubrics contain a description of the four performance levels with examples of possible answers for each performance level alongside the initial description. The final teacher evaluation result is an average of the four assessment instruments. Scores are weighted by their relative importance (Self-evaluation: 10%, Peer evaluation: 20%, Third-party assessment report: 10% and Portfolio: 60%). The overall results classify teachers into one of the four performance levels previously mentioned (Flotts & Abarzua, 2011).

Teachers who are rated unsatisfactory have to be reevaluated the following year. If the teachers are evaluated as unsatisfactory in two consecutive evaluations, they have to leave the municipal educational system. With respect to the teachers who are evaluated as basic, two years later they are required to be reevaluated[1]. In addition, they must participate in Professional Development Plans specifically designed to address their

[1] After the Law 20,501 in 2011 was passed, teachers have been categorized as "basic" three consecutive times are asked to leave the municipal system.

weaknesses identified in the evaluation (Cortes & Lagos, 2011). Meanwhile, those teachers who are evaluated as outstanding or competent only have to be reevaluated every four years, and they can choose to be part of the Variable Individual Performance Allowance program (Asignación Variable por Desempeño Individual [AVDI], in Spanish[2]), which was a voluntary reward program. This program includes a standardized test that assesses the disciplinary and pedagogical knowledge of teachers and includes monetary rewards based on the results. Between 2003 and 2017, more than two hundred thousand teachers participated in this process, including primary, secondary, technical professional, and early childhood teachers (Docentemas, n.d.).

### Teacher Professional Development System.

In 2015, during the second governance term of the former Chilean president Michelle Bachelet (2014-2018), an educational reform made up of a series of initiatives and bills whose purpose was to produce deep transformations in the Chilean educational system was implemented. These initiatives were sought to guarantee the right for every student to quality education from their first years until they graduated from higher education (Centro de Estudios, 2017). In the context of the Educational Reform, the new Teacher Professional Development System was put into low in 2016. Rather than replacing the Chilean National Teacher Evaluation System, the new system complement it. The goal of the new system is to improve the quality of initial teacher preparation, coursework, and teaching practices. The system commits to supporting teachers from the

[2] AVDI (Variable Individual Performance Allowance) was a voluntary annual reward program between 2004 and 2015. Results of the AVDI program were used to award bonuses to teachers who succeeded in their application.

beginning to the end of their careers, which implies continued professional development (OECD, 2017). In addition, the new system incorporates all teachers and educators who work in nationally funded schools and preschools (municipal schools, charter schools, and schools with delegated administration).

The system will be implemented progressively between the years 2016 to 2026 for all teachers who work in nationally funded schools. As part of the Teacher Professional Development System, all teachers working in nationally funded schools are currently categorized into three levels: Initial, Early, and Advanced teaching. Additionally, there are two voluntary levels that teachers can opt to be a part of, called Expert I and Expert II. The teacher's categorization and progression through these levels are based on the following three factors: their years of experience, the results of the **teacher performance portfolio** in their last evaluation[3], and the results from the **standardized test of disciplinary and pedagogical knowledge** (Ruffinelli, 2016). The system develops a career and new pay structure for teachers and promises to increase the value of the role of teachers and the teaching profession in the community. Therefore, advancement to the next level means the possibility of taking on new responsibilities and receiving a higher salary (OECD, 2017).

The following figure shows the relationship between both systems that are currently used in Chile.

[3] The same portfolio used in the Chilean National Teacher Evaluation System is now used by the System of Teacher Education and Professional Development to categorize the teachers.

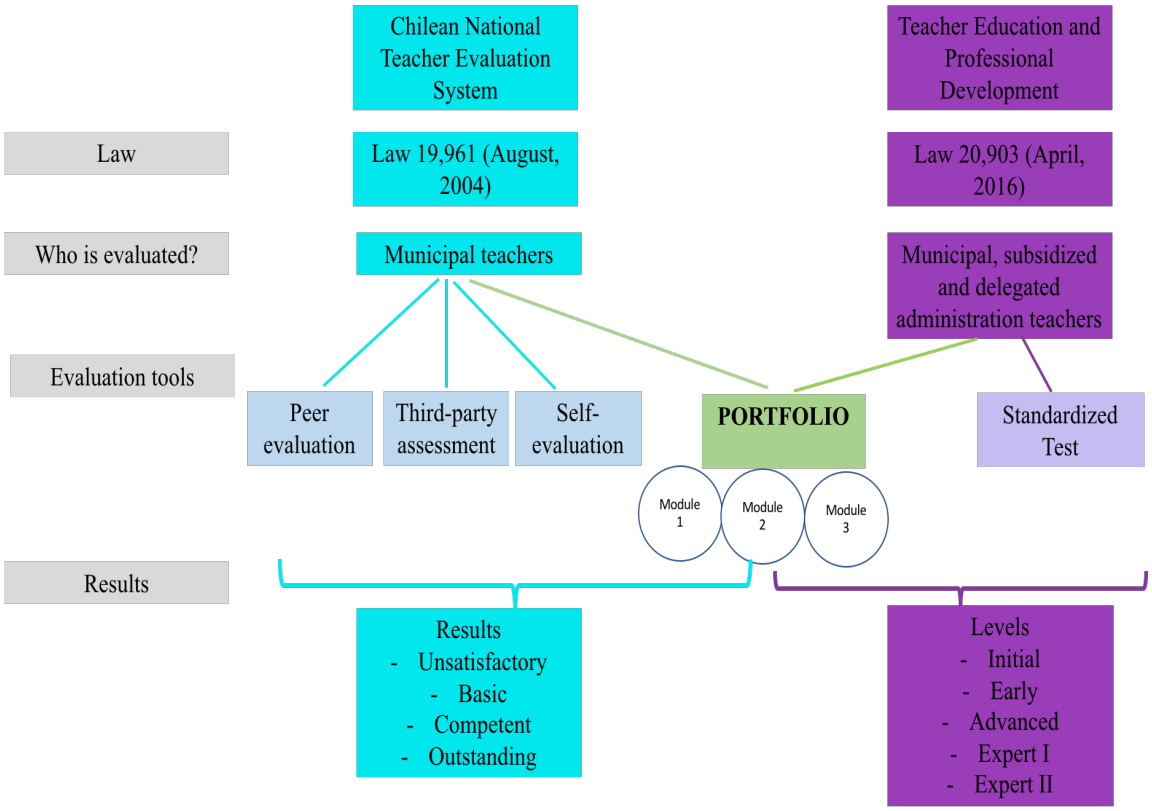


*Figure 1.1.* Relationship between Chilean National Teacher Evaluation System and the Teacher Education and Professional Development. Translated and adapted from www.docentemas.

## Portfolio Assessment Instrument

A portfolio assessment instrument is a tool used to evaluate students, teachers, schools, and so on. In the teachers' case, the portfolio normally consists of various pieces of evidence, such as lesson planning, students' notebooks, and class records. The person evaluated selects material to be assessed as a part of the portfolio, and other people (school authorities, peers, or external evaluators) evaluate that material. Research indicates that teacher portfolios can be as useful in learning about a particular teacher's practice as classroom observation (Martínez et al., 2012).

International evidence shows various examples of portfolio use as a measurement tool for teacher quality. For instance, Singapore includes an electronic portfolio in the teacher preparation program, which contains a collection of evidence that documents student teachers' growth in their capacities and abilities over the course of their teacher training (The National Center on Education and the Economy, 2016). The Performance Assessments (PPAT), developed by the Educational Testing Service (ETS) in the USA, includes an evidence-based portfolio assessment designed to promote learning and to guide teaching candidates' practices (Educational Testing Service, 2018). The EdTPA Support and Assessment Program, created by Stanford University's faculty and staff, is an authentic assessment tool that aims to describe and document practices in a portfolio. This system has been adopted by 25 states in the United States (Stanford Center for Assessment, Learning, and Equity, 2013). Another example is the portfolio that has been used in the Chilean National Teacher Evaluation System and the Chilean Teacher Professional Development System (Taut et al., 2012).

## Chilean Portfolio Assessment Instrument

As it was possible to observe in Figure 1.1, the portfolio is the only assessment instrument shared by both the National Evaluation and the Teacher Professional Development System. It should be considered to be the core evaluation tool, as it is the most complex part of the evidence that the teachers have to submit, and contributes the most to the calculation of the overall score for the evaluation for the Chilean National Evaluation System (60%).

The portfolio has been designed for teachers to provide evidence of their best pedagogical practices (Santiago & OECD, 2013). The GTF proposes different domains, broken down into indicators, to be evaluated within the portfolio. Thus, each dimension is rated by indicators in the portfolio, that are grouped into three modules (Flotts & Abarzua, 2011):

a) Module 1: Set of pedagogical materials that require the teacher to plan and implement an 8-hour teaching unit, design an end of term assessment for the teaching unit, and respond to a set of questions about teaching practices that evaluates how the teacher incorporates the characteristics of their group of students when planning the teaching unit. For Module 1, each teacher can choose to be evaluated on one out of two teaching units within their subject and grade level with predetermined learning objectives from the Chilean National Curriculum.

b) Module 2: Video recording of a class, which consists of a 40-minute recording of a regular class along with a questionnaire about the class. The teacher knows

ahead of time when he or she will be videotaped. The recording of the class is independent of the other tasks requested by the portfolio; thus, it is not related to the unit designed as part of Module 1.

c) Module 3: This module is not mandatory and it was included in the portfolio after 2017. The module measures proof of collaborative work understood as the interaction, experience exchange, and shared reflection between the teacher evaluated and their school peers. In this module, the principal must also provide reports related to the teacher's professional development and the additional responsibilities accomplished by the teacher in the school.

Portfolio modules and indicators have been changing gradually since the implementation of the system. One of the biggest changes was the inclusion of a third module in 2016, as part of the Teacher Professional Development System Law. This module incorporated new dimensions of teaching work, such as collaborative work with other members of the educational community, professional development, and other school responsibilities besides classroom work. In the portfolio final score for both the National Teacher Evaluation System and the Teacher Education and Professional Development System, the results of the indicators that Module 3 evaluates are only taken into account when it benefits the teacher's results (docentemas.cl).

Also, for both systems, there are four aspects evaluated within the portfolio that have a greater weight when calculating the final score of the instrument: curricular emphasis on the subject, clear explanations, questions and activities, and student feedback and use of assessment results. This difference is justified because those are specific aspects of the pedagogical interaction that is established between the teacher and

their students, considered as essential elements of effective teaching, and that has a direct impact on their learning (Darling-Hammond, 2000; Donker et al., 2014; Hattie & Timperley, 2007).

Table 1.2 shows modules, dimensions, indicators, and weights for the teacher evaluation portfolio:

Table 1.2.
*Portfolio Modules, Dimensions, Indicators, and Weights for each indicator.*

| Modules | Dimensions | Indicators | %[4] | %[5] |
|---|---|---|---|---|
| Module 1: Set of pedagogical materials | 1. Planning[6] | 1.1. Formation of learning objectives | 5% | 4% |
| | | 1.2. Relationship between activities and objectives | 5% | 4% |
| | 2. Assessment | 2.1. Evaluation and rubrics used for correction | 5% | 4% |
| | | 2.2. Relationship between assessment and objectives | 5% | 4% |
| | | 2.3. Analysis and use of assessment results | 5% | 4% |
| | 3. Reflection | 3.1. Analysis based on students' characteristics | 5% | 4% |
| | | 3.2. Use of error for learning | 5% | 4% |
| Module 2: Video recording of a class[7] | | 4.1. Class environment | 5% | 4% |
| | | 4.2. Quality of the start of class | 5% | 4% |
| | | 4.3. Quality of the end of class | 5% | 4% |

[4] These percentages refer to the weight of each indicator when Module 3 is not included in the total score.
[5] These percentages refer to the weight of each indicator when Module 3 is included in the total score.
[6] For Technical Education Teachers there is one more indicator for this dimension: Integration of general learning in activities.
[7] For Technical Education Teachers the indicators 4.2 and 4.3. are not evaluated. Instead, there are two more indicators for this dimension: Activity Monitoring and Link to the Working World.

| Modules | Dimensions | Indicators | %[4] | %[5] |
|---|---|---|---|---|
| | | 4.4. Contribution of the activities to the achievement of the class objectives | 5% | 4% |
| | | 4.5. Curricular emphasis on the subject | 10% | 9% |
| | | 4.6. Clear explanations | 10% | 9% |
| | | 4.7. Questions and activities | 10% | 9% |
| | | 4.8. Encouragement | 5% | 4% |
| | | 4.9. Student feedback and use of assessment results | 10% | 9% |
| Module 3: Collaborative work | | 5.1. Collaborative work suitability | | 4% |
| | | 5.2. Quality of professional dialogue | | 4% |
| | | 5.3. Value of collaborative work in professional development | | 4% |
| | | 5.4. Reflection on the impact of the collaborative work experience | | 4% |

*Note.* Translated and adapted from www.docentemas.cl.

The portfolio evaluation is assessed in correction centers[8], located at different universities across the country, and whose selection requires the approval of the Ministry of Education. MIDE center of Pontifical Catholic University of Chile (MIDE UC) is responsible for the design and planning of the correction process, as well as for the supervision (in person and remotely) of each correction center. In addition, MIDE UC is responsible for distributing the portfolios to the centers, and considering the grade level and subject that the center evaluates. Each correction center requires certain infrastructure

[8] More detailed information for the Portfolio correction is presented in Section 3.

and the necessary equipment, as well as a personnel selection process to hire qualified correctors and correction supervisors. Classroom teachers, who are specially selected and trained to perform this task, correct the portfolios. Both supervisors and correctors must have training and experience at the same grade level and subject area as the teachers who are being evaluated. The portfolio correction follows a series of strict quality procedures using standards that ensure a quality evaluation process (Sun et al., 2011).

## Evaluation Results Reports

The evaluation process for both the Chilean National Teacher Evaluation System and the Teacher Professional Development System in Chile ends with the results reports, which are submitted to each participating teacher. Through the report, each teacher is informed of their strengths and weaknesses in their teacher performance (Sun et al., 2011). Each teacher receives an Individual Results Report, which is a confidential document that reports the results of the four evaluation instruments and provides detailed feedback on the strengths and weaknesses in the portfolio. Furthermore, the report provides the teacher with their overall performance level according to the Chilean National Teacher Evaluation System. Also, the teachers receive the results of each indicator evaluated in the portfolio, designating them to one of four performance levels: unsatisfactory, basic, competent and outstanding (Docentemas, 2019). This feedback is considered an essential part of the process that requires continuous reflection and review to serve the purpose of promoting development and good performance (Santiago & OECD, 2013).

The results for the Teacher Evaluation System are also submitted in different ways to other parties involved: local school boards (municipalities) and principals. For

the schools and municipalities, the report gives them aggregate information about the teachers' performance, which provides a general overview of their teachers, as well as helping them to pinpoint the professional development areas needed. The purpose of the report for the teachers is to show them their strengths and weaknesses to incentivize their professional growth (Docentemas, 2019).

Each teacher receives an Individual Results Report, which is a confidential document that reports the results of the four evaluation instruments and provides detailed feedback on the strengths and weaknesses in the portfolio. Furthermore, the report provides the teacher with their overall performance level according to the Chilean National Teacher Evaluation System. Also, the teachers receive the results of each indicator evaluated in the portfolio, designating them to one of four performance levels: unsatisfactory, basic, competent and outstanding. The Individual Results Report also provides a final score for the portfolio that is used to categorize the teacher in the Teacher Professional Development System (Docentemas, 2019).

## Purpose of the Study

The purpose of the present dissertation is to contribute to the body of research on the evidence of the validity of Chilean Teacher Evaluation System, focusing specifically on the portfolio. From the present research relevant evidence of internal structure validity of the portfolio will be presented, contributing to the new agenda of teacher evaluation validity carried out by the MIDE UC Research team.

Validity has been defined by the *Standards for Educational and Psychological Testing* as the degree to which evidence and theory support the interpretations of test scores for the proposed uses of tests (AERA et al., 2014). This definition is consistent

with the argument-based approach, which indicates that "to validate an interpretation or use of measurements is to evaluate the rationale, or argument, for the proposed conclusions and decisions" (Kane, 2006, p. 17). The validation process calls for "a clear statement of the proposed interpretations and uses and a critical evaluation of these interpretations and uses" (Kane, 2006, p. 17).

This approach is usually used in validating educational testing programs, like the Chilean Teacher Evaluation System. Therefore, in order to contribute to the evidence of the validity of the Chilean Teacher Evaluation System, one important step is to clarify how the test scores will be interpreted and the purpose for which they will be used (Taut et al., 2012).

For the Chilean Teacher Evaluation System, two broad purposes have been recognized by the stakeholders: providing formative data on individual teachers to improve their practice, and providing summative data to support individual teacher rewards and sanctions (Taut et al., 2012). The previous research for the system validation took into account these two evaluations aims, contributing to the Chilean Teacher Evaluation System with validity evidence in making the system more valid and relevant for all assessment users.

The present research intends be an extension of previous research carried out by the researchers from MIDE UC, taking into consideration that from 2016 on, the portfolio had important modifications such as the inclusion of a completely new module that evaluates collaborative work (Module 3). This research is also taking into consideration the formative and summative purposes of the Chilean Teacher Evaluation System. Therefore, evidence of validity for the portfolio in the present research will focus first on

the summative purpose of the evaluation. Taking this purpose into consideration, first I will provide evidence for the portfolio correctly distinguishing overall teacher quality for teachers who teach in different contexts or settings. Second, I will provide evidence for the structure validity considering the portfolio's new structure. Third, I will provide evidence for the portfolio final score based on weighted indicators, comparing them with the empirical evidence.

A second portfolio purpose is their formative intention. Then, evidence of portfolio validity based on the formative purpose will focus on the identification of teacher strengths and weaknesses using validity of portfolio subscores.

Consequently, the specific aims of this study are:

**Aim 1**: Assess the structure of the portfolio (used in the Chilean Teacher Evaluation System after 2016) across two different subgroups: teachers whose Module 3 evaluation was taken into account for their final portfolio score, and those teachers whose Module 3 was not taken into account for their final score

**Aim 2**: Determine if the portfolio factor structures are invariant across subgroups such as different teaching levels and school location (rural/urban).

**Aim 3**: Compare the theoretical weight assigned to each one of the portfolio indicators with the empirical data from the Chilean Teacher Evaluation System.

**Aim 4**: Evaluate validity evidence that supports the interpretation and use of portfolio subscores.

This aim will be accomplished first by using exploratory Confirmatory Factor Analysis (ECFA) to evaluate the portfolio's new structure taking into consideration the inclusion of Module 3. Through this analysis, I will determine the number of factors

underlying the **20 portfolio indicators** for the whole sample of teachers evaluated and for the subgroup of teachers whose Module 3 score was taken into account for their final portfolio score, and to determine the factors for the **16 portfolio indicators** for the subgroup of teachers whose Module 3 score was not taken into account for their final score. The ECFA approach incorporates the EFA and CFA models, allowing more accuracy since it avoids potential pitfalls due to the challenging EFA to CFA conversion by estimating the measurement and structure model parts simultaneously (Asparouhov & Muthén, 2009).

Second, I will also provide validity evidence for the portfolio factor structures by focusing on teachers that teach in different contexts or settings. I will explore the evidence for multigroup invariance within identifiable subgroups of teachers working in different school contexts. It is generally recognized that the context affects the evaluation of teaching performance. The quality of teaching - the results of the educational process - is strongly influenced by the instructional context (Bryk et al., 2012; Darling-Hammond, 2010), therefore teachers should be evaluated based on the institution, the student population and the resources with which they work (Everson et al., 2013).

For instance, if adverse characteristics impede teachers' performance, this context will be classified as difficult. This should be the case for Chilean rural teachers, who in addition to working in multi-grade classrooms, mostly must work in areas of difficult access, vulnerability, and high poverty rates. Considering that context, teachers from rural areas face the teacher evaluation process on a different footing than their urban pairs; that difference is not necessarily considered for the evaluation system (Castillo-Miranda et al., 2017). The rural sector is at a disadvantage and put under unequal

conditions in the Chilean Teacher Evaluation System. In general, rural teachers do not come out well evaluated due to the conditions that exist, for example they often have no internet connection. Teachers must answer the portfolio online, so there is a connectivity problem that affects their result (Colegio de Profesores de Chile A.G., 2016).

There is also evidence from the earliest generation of observation instruments and teacher surveys that teachers practice differently across grade levels (Vartuli, 1999). Research has shown that elementary school teachers obtain better rating on their evaluations from principals than middle school teachers (Harris & Sass, 2011). Therefore, teacher grade level could impact the portfolio teacher result.

This aim intends to contribute to the study of context differences, exploring validity evidence of portfolio factorial invariance, considering rural and urban context, and teaching levels, taking into consideration that the same portfolio is used by a wide range of teachers from very different school contexts. Factorial invariance analysis is an important procedure in studies that seek to make comparisons between two or more groups that use the same evaluation instruments, certifying that the structural features of the instrument remain unchanged between the compared groups (Byrne, 2008). Thus, when considering different samples of teachers, an item or indicator possesses the same amount of difficulty across samples (Hambleton et al., 1991)

Third, I will also provide validity evidence for the portfolio results based on the weighted score assigned to each indicator that was also introduced with the new portfolio structure. Previous portfolio validity research indicated that the portfolio contained a number of different elements, but the final score does not indicate how the different "pieces or indicators" will be combined together to create a single "score" (Santiago &

OECD, 2013). Therefore, validity evidence related to the way that different portfolio indicators are combined to composite a unique score will be tested using a weighted sum score method. The weighted sum score method is that items with the highest loading on the factor have the biggest effect on the factor score (Distefano et al., 2009).

Finally, validity evidence for the portfolio formative purpose will be evaluated taking into consideration the inclusion of more detailed information on possible portfolio subscores. The portfolio contains separate sections or domains of evidence that have to be submitted, therefore possible information related to each domain can be reported if each subscore contributes to more accurate measurements of the construct than is provided by the total score (Haberman, 2008). In order to evaluate the use of subscores, one of the approaches used by Haberman (2008) is the **proportional reduction of the mean squared error (PRMSE).** For the present dissertation I use Feinberg and Wainer (2014), who refined the Haberman (2008) method, by presenting the *PRMSEs* as a ratio that they called value-added ratio (VAR). Through the calculation of value-added ratio for each possible portfolio subscore, results greater than one indicate that that subscore shows added value over the total score, therefore serving as a plausible teacher evaluation result.

Based on the findings of the portfolio's validity evidence from this research, this study examines the option to improve this tool, including modifications in the rubric used to correct the portfolio, as well as changes in the final report given to teachers, administrators, and employers.

## Significance of the Study

This study will be a contribution to the validity plan for the Chilean National Teacher Evaluation System and the Teacher Professional Development System, especially considering that the portfolio from the teacher evaluation system is intended to be used for high-stakes decisions. The results of the portfolio will now consider the new career structure, with a salary scale associated with each career level. Therefore, any research that contributes to the validation of the system is essential. The *Standards for Educational and Psychological Testing* (AERA et al., 2014) suggest that we have a professional responsibility to engage with and monitor the validity evidence for any large-scale testing and examination system.

The importance of focusing on the portfolio is based on the fact that of the four instruments used in the Chilean Teaching Assessment System, the portfolio is the only one that provides direct information on teaching work and it has the most weight in the calculation of the overall score for the evaluation (Sun, Calderon, Valerio, & Torres, 2011). Also, the portfolio is the only instrument that overlaps on both the Chilean Teaching Assessment System and the new Teacher Professional Development System. The portfolio is the most complex part of the evidence that the teachers have to submit, , and presents the highest correlation to the student results measured by the Chilean National Standardized Test (SIMCE) (Alvarado et al., 2011).

Considering the high-stakes consequences of the results of the instruments applied in the Teacher Professional Development System, and the fact that the portfolio is the core instrument in the evaluation, a detailed and permanent investigation of its validity is necessary. This research attempts to enrich this objective.

## Limitations of the Study

The database used is from 2017. Therefore, it includes information from the first year in which the Teacher Professional Development System was implemented. Since then, the system has been implemented gradually, with the intention of continued gradual change until the year 2026. One limitation is that because 2017 was the first year of the new evaluation system, it was not possible to have data for teachers who work at charter schools. Thus, analysis such as invariant factor structures across municipal and charter school teachers, was not possible. Second, this dissertation focuses on one type of validity evidence described by the *Standards*, which is internal structural validity. Ideally, there would be other parallel researchers interested in complementing the validity of the portfolio. Third, the current study focuses on only one the instruments that are being used as part of the Teacher Evaluation System and the Teacher Professional Development System. Future research should concentrate on the other evaluations tools, the Standardized Test used to measure disciplinary and pedagogical knowledge, or on how to combine the data from both of those different evaluations.

## Summary

The introductory chapter describes the overall goal of this dissertation, which is to contribute to the body of research on the evidence of the validity of the Chilean National Teacher Evaluation System and Teacher Professional Development System, focusing specifically on the portfolio. I also provide some background information on the Chilean Educational System, the development of the Chilean National Teacher Evaluation System, and Teacher Professional Development System. Later on, I focus on the portfolio

evaluation instrument, describing the different domains and indicators evaluated. Finally, I present the purpose of the present study, and each specific aim that will incorporate relevant evidence of portfolio validity.

## Organization of the Study

Chapter Two describes a revision of the literature of teacher quality conceptualization, frameworks that have been developed in this context, and different teacher evaluation systems. Further, I review the conceptualization of validity. I review sources of validity of teacher quality evaluation systems, including the prior validity studies for the Chilean Evaluation Systems. Chapter Three describes the data source, the teacher evaluation measures used in the Chilean system, and the analytic approach that will be used to support the aims. Chapter Four presents the results of the aims. Chapter Five concludes with a discussion of the results, as well as the recommendations, limitations, and future studies of teacher evaluation system validity.

# Chapter 2. Literature Review

The core of most high performing education systems is effective evaluation of teachers' performance on a regular basis. Teacher evaluation plays a critical role in supporting and improving the quality of teachers and in holding them accountable. Therefore, it plays a formative role since teacher evaluation identifies overall weaknesses, and an accountability role can be used as a platform for rewarding high performers (Bruns & Luque, 2014). Evidence shows that investing in well-designed systems to evaluate teachers at regular intervals and give them timely feedback on their practices contributes to their effectiveness, leading to positive benefits for their students (Taylor & Tyler, 2012).

Bruns and Luque (2014) define key steps in the design of a teacher evaluation system. One key step refers to the definition of good teaching, which implies the creation of national teaching standards. Also, a key step is the identification of the mechanism with which teaching can be measured, including the development of instruments that can produce valid estimates of teachers' effectiveness. In this literature review chapter, I focus on the definition of teacher quality from different approaches and frameworks. I present different teacher evaluation systems around the world, identifying the purposes, standards, and frameworks at the base of the system, and the instruments used to evaluate teachers. Later, I focus on the concept of validity and how different sources of validity have been evaluated in teacher evaluation systems in both the United States and Chile.

Finally, I present limitations of the literature review and the contribution of the present dissertation to fill in those gaps.

## Teacher Quality

The term teacher quality is going to be used in the present literature revision to refer to the characteristics of individual teachers and the processes and practice that teachers employ in their teaching (Cortez-Ochoa et al., 2018). Teacher effectiveness, as measured by student outcomes, is not going to be considered in the review of teacher quality because the mechanisms for linking student outcomes to individual teachers are not part of the Chilean Teacher Evaluation System, which is the focus of the analysis in the present dissertation.

A fair and valid teacher evaluation system needs criteria and standards to define what is understood as teacher quality (Isoré, 2009). Standards refer to statements describing what is expected of a teacher's knowledge and performance in their day-to-day teaching, developed as guidance for making judgments about those teachers (Cortez-Ochoa et al., 2018). Professional standards for teachers make the expectations explicit for what a teacher's role entails. Given that there are different teaching abilities and responsibilities, professional standards can provide systematic coherence and clarify what the country considers to be a good teacher (OECD, 2017; Sartain et al., 2011). Moreover, defining a good teacher can serve as the basis for the formulation of the different elements included in the standards (Darling-Hammond, 2010; Schmelkes, 2015).

Additionally, frameworks have been defined as the mechanisms used in order to reach judgments about whether a teacher has attained the standards or not. Frameworks may include sets of standards for teachers at different stages of their careers, but they also

set out the pathway for moving from one stage to the next. Therefore, based on the standards, frameworks for teaching can be designed, providing methods to evaluate teachers' performance by describing different levels of achievement for each component of the framework (Cortez-Ochoa et al., 2018).

There are many frameworks that include different standards of what would be conceptualized as a good teacher and that have been used from a range of international contexts. In the present literature review, the frameworks presented were selected based on their support by research, and because they are widely used and adapted in different settings (Bill & Melinda Gates Foundation, 2012; Clinton et al., 2017; OECD, 2009).

**Charlotte Danielson's Framework for Teaching** is one of the most well-known models and a reference point in this field (Goe, 2007; Taylor & Tyler, 2012), and various national and local teacher evaluation systems have been influenced by this framework (OECD, 2013). Danielson's framework is a multifaceted research-based conception of teaching, describing what teachers do in their professional practice (T. J. Kane et al., 2014), and identifies those aspects of a teacher's responsibilities that have been documented as promoting improved student learning through empirical and theoretical research (Danielson, 2007). This framework describes 22 components that outline teacher responsibilities that are then clustered into the following four domains, that refers to different aspects of teaching:

1) **Planning and Preparation**, describe how a teacher organizes the content that the students are supposed to learn. This domain covers all aspects of instructional planning, beginning with a deep understanding of content, pedagogy, and the appreciation of the students and what they bring to the educational encounter.

2) **The Classroom Environment**, establishes a comfortable and respectful classroom environment that cultivates a culture of learning and creates a safe place for risk-taking.

3) **Instruction,** represents the implementation of the plans designed in Domain 1. Teachers demonstrate, through their instructional skills, that they can successfully implement those plans. Includes a wide range of instructional strategies that enable students to learn.

4) **Professional Responsibilities**, associated with being a true professional educator, going beyond the classroom responsibilities, connecting with the students outside of the classroom. Includes self-assessment, communication with parents, participating in ongoing professional development, and contributing to the school and district environment. Even though Domain 4 relates to professional responsibilities, it encompasses the roles assumed by the teacher outside of the classroom, in addition to those in the classroom, with the students. It takes into consideration activities that are critical to preserving and enhancing the teaching profession. Educators exercise their professional responsibilities because they are integral to the work they do with their students. The inclusion of Domain 4 is one of the contributions of Danielson's framework, since it is through the skills evaluated in this domain that highly professional teachers distinguish themselves from their less proficient colleagues (Danielson, 2007).

Each one of the components defines a distinct aspect of a domain, and it is associated with levels of performance that range from describing teachers who are still striving to master the rudiments of teaching to highly accomplished professionals who are able to share their expertise. The levels of performance are: **Unsatisfactory**, a teacher that does not yet appear to understand the concepts underlying the component; **Basic**, a

teacher that appears to understand the concepts underlying the component and attempts to implement its elements. However, implementation is sporadic, intermittent, or otherwise not entirely successful; **Proficient**, a teacher that clearly understands the concepts underlying the component and implements them well; **Distinguished**, a teacher who demonstrates mastery, making a contribution to the field, both in and outside of their school. Their classrooms operate at a qualitatively different level from those of other teachers (Danielson, 2007).

One strength of Danielson's points regarding this framework is the fact that it could be used in a generic way; that means that it could be used for generalist teachers, as well as for subject specific educators, from different contexts and situations (Danielson, 2007). Additionally, the levels of performance described in the framework are especially useful in supervision and evaluation. However, they can also be utilized to help with self-assessment or to support mentoring, by generating a professional discussion and suggesting areas for further growth. Therefore, Danielson's Framework for Teaching can serve both summative and formative purposes (Isoré, 2009).

Charlotte Danielson's Framework for Teaching has been very influential in the United States context, adopted as part of evaluation systems in different states (Cortez-Ochoa et al., 2018; Lazarev et al., 2014). For instance, the **Teacher Advancement Program (TAP)**, operated by the California-based National Institute for Excellence in Teaching, uses a set of standards for evaluating teachers based on the work of Danielson. TAP's modified version of Danielson's teaching standards has three main categories: designing and planning instruction, the learning environment, and instructions (Toch, 2008). Also, Danielson's Framework has influenced the evaluation systems overseas in

countries such as Chile, Peru, and Mexico states (Cortez-Ochoa et al., 2018; Taut & Sun, 2014; Vázquez Cruz et al., 2014).

A similar framework to Danielson's for teacher's performance evaluation has been developed by James Stronge. He created the **Goals and Roles Performance Evaluation Model** with the aim to improve student learning and teachers' practices by collecting evidence and presenting data to document teachers' performance based on well-defined job expectations (Stronge, 2012). This model clearly defines professional responsibilities, consisting of six performance standards, that refer to the major duties performed, and a flexible number of performance indicators, that provide examples of observables and tangible behaviors (Stronge, 2010).

The six performance standards are:

1) **Instructional Planning**, related to the teacher's plans using the school's curriculum, effective strategies, resources, and data in order to meet the needs of all their students.

2) **Instructional Delivery**, described as how the teacher effectively engages students in learning by using a diversity of instructional strategies to meet individual students' learning needs.

3) **Assessment of/for learning**, how the teacher systematically gathers, analyzes, and uses data to measure their students' progress and to guide instruction in order to provide timely feedback.

4) **Learning Environment**, the use of resources, routines, and procedures by the teacher in order to provide a respectful, positive, safe, and student centered environment that is conducive to student learning.

5) **Professionalism**, in which the teacher maintains a commitment to professional ethics, international mindedness, and the school's mission. Also, professionalism indicates that the teacher takes responsibility for and participates in professional growth that results in the enhancement of student learning.

6) **Student Progress**, refers to the association between the work of the teacher as a result of acceptable and measurable students' progress (Stronge, 2010).

For this framework, a fair and equitable evaluation system for the teacher's performance necessarily requires the collection of multiple data sources in order to provide a comprehensive and authentic performance portrait of the teacher's work. The sources of information proposed are: goal setting for student progress, observations, teacher documentation folder, and student surveys (Stronge, 2010).

The rating scales for each evaluation instrument are put into four levels of how well the standards are performed on a continuum from exemplary to unacceptable. **Exemplary** refers to a teacher who exceeds the expectations. Those who meet the standards are **proficient**. The two lower levels: **developing/need improvement** and **unacceptable** are described for teachers who do not meet the expectations (Stronge, 2010).

Another commonly used teacher framework across the United States, Canada, and Australia, as well as countries within Europe, Asia, and South America is the **Marzano Teacher Evaluation Model**. This model was designed for formative uses, and as part of a supervision-based strategy to improve teachers' instructional skills. It is a scientific-behavioral evaluation system designed and created by using an aggregation of the

research on elements traditionally shown to correlate with student academic achievement (Clinton et al., 2017).

The system foments reliability for observers and simplifies the evaluation process, emphasizing observable elements with specific evidence of teacher effectiveness. The model identifies key elements, or professional and instructional strategies, divided into four domains, designed to progressively guide a teacher from planning, to implementation of instructional strategies, to awareness of conditions for learning in the classroom, and lastly to professional responsibilities (Carbaugh et al., 2017).

Thus, the model concentrates measurable teacher actions and capabilities into 60 elements that evaluate measurable behaviors of effective teachers to be scored, within the four domains:

1) **Classroom Strategies and Behaviors** (41 elements), that clearly emphasizes what occurs in the classroom.

2) **Planning and Preparing** (8 elements), both of which are assumed to be directly linked to classroom strategies and behaviors. Careful planning and preparation give a teacher enough time to incorporate effective classroom strategies and behaviors.

3) **Reflecting on Teaching** (5 elements), that focuses on self-reflection, which has been considered as a vital metacognitive step in teacher development.

4) **Collegiality and Professionalism** (6 elements), that focuses on teacher professional behavior, which is only indirectly linked to classroom strategies but it makes up the foundational expertise from which the preceding three domains can flourish (Marzano & Toth, 2013).

The model utilizes a common five-point scale that provides a developmental continuum for teachers based on five levels of proficiency: **Not Using**, the strategy evaluated was called for but not exhibited; **Beginning**, the strategy is used incorrectly or with parts missing; **Developing**, uses the progression of standard-based learning targets embedded within a performance scale to identify accurate critical content during a lesson, but less than the majority of students are displaying the desired effect; **Applying**, uses the progression standard-based learning targets embedded within a performance scale to identify accurate critical content during a lesson, and the desired effect is displayed in the majority of students; **Innovating**, based on student evidence, implements adaptations to achieve the desired effect in more than 90% of the student evidence (Carbaugh et al., 2017).

**Ronald Ferguson's 7Cs framework** from **The Tripod Project** survey assessment has also been a popular way for measuring what teachers actually do in their classroom and to diagnosing teachers' professional strengths along with areas in need of improvement, in the United States. As the other approaches reviewed in this chapter, the 7Cs framework is research-based and has been refined based on analyses of prior results and feedback from school practitioners and fellow researchers (Ferguson & Danielson, 2015).

The 7Cs framework is derived from peer-reviewed research published in education books and journals. It is grouped into seven scales that measure teacher quality: 1) **Care**, teachers that show concern and commitment. They develop supportive, personalized relationships with students and strive to cultivate an emotionally safe environment where all students feel respected and learning is the central focus.

2) **Confer**, teachers that promote ideas and discussion. These teachers seek and value students' points of view, providing frequent opportunities for students to share their perspectives.

3) **Captivate**, teachers who inspire curiosity and interest. They make instructions engaging, with lessons that are frequently intriguing and relevant to students and hold students' attention.

4) **Clarify**, teachers who frequently check for understanding, address misconceptions, explain ideas and concepts in a variety of ways, and provide useful feedback.

5) **Consolidate**, teachers who help students organize content in ways that make it easier for them to remember and reason efficiently. They summarize the learning at the end of each lesson, highlighting relationships between ideas.

6) **Challenge**, teachers that are concerned with persistence and rigor. They hold students' to high academic and behavioral standards and monitor student's effort.

7) **Control**, teachers who vigilantly monitor students' behavior, manage and redirect off-task behaviors, and foster classroom conditions that allow for optimum learning (Tripod Education Partner, 2014).

Each one of the seven components is measured by multiple items through a student survey. They are surveyed for different grades, however all the versions cover the same concepts evaluated, although some items are worded more simply for the elementary school version (Ferguson & Danielson, 2015).

Although the frameworks described previously are not exhaustive, it is illustrative of the variety of options currently used in various countries. Also, they share common elements based on the literature, experts' opinions, and empirical evidence. The Chilean

Teachers Evaluation System, took as its foundation most of the characteristics related to quality teachers presented and shared by the frameworks described previously, indicating that it is a system of evaluation based on the evidence from research.

## Teacher Evaluation Systems

As was previously discussed, teacher evaluation has been put forward as an important strategy for assuring and developing educational quality worldwide (Skedsmo & Huber, 2018). Around the world, different countries have reported and implemented well-designed practices to provide feedback on teachers' work (Huber & Skedsmo, 2016). However, each system presents radical differences in terms of the purposes, approaches, and instruments used for the evaluation, among others.

In this section, different teacher evaluation systems will be reviewed. Some of them were selected because they are considered top performing OECD countries. Although Chile is not considered a high-income country, it was the first country in South America to join the OECD in 2010. This implies that for the country participants in all of the organizational areas, including education. In this way, Chile joins with other countries to share experiences, but also to set new standards, aiming at the countries that make up the organization (OECD, 2013). Additionally, two countries from the region that have been mentioned very often in the literature for the implementation of their national teacher evaluation system (Vaillant, 2008), were also included in the present review.

In the United States, most states have been incorporating teacher evaluation practices. Teacher evaluation was embedded into state law and school district practice in 2002 with the No Child Left Behind (NCLB) Act. To qualify for a NCLB waiver, states were required to develop evaluation systems with continuing educator input, clear and

useful feedback, use of multiple measures that could include student growth, differentiated teacher performance, and informed personal decisions (Aragon, 2018). However, the way that it has been implemented differs depending on the state. In most states, teacher evaluation is based on students' tests scores (value-added models), adding in some cases the development of a portfolio or class observation as part of the teacher certification process (Vaillant, 2008). Further, the Every Student Succeeds Act (ESSA) from 2015, provided new flexibility to states to revise and reform their teacher evaluation systems, and states now have full discretion over whether and how to evaluate teachers (Aragon, 2018).

In general, for all the states, the goal of teacher evaluation is to collect data that accurately represents teacher practices and use that information to improve the system (Cleaver et al., 2018). Accurate evaluation can help differentiate teacher performance, inform feedback, improve professional development, and provide opportunities for advancement or rationale for teacher dismissal (Aragon, 2018).

In terms of instruments used to evaluate teachers in the United States, students' test scores data have been incorporated into the state teacher evaluation systems as a measure of achievement or mastery (Cleaver et al., 2018). Currently, 34 states require teacher evaluations to include measures of student growth. However, approximately one-quarter of the 34 states do not currently require the state's standardized test to be the source of those data. They have been shifted to the use of measures such as district assessments, student portfolios, and student learning objectives to determine teacher's contributions to student growth. Observations by school leaders, administrators, or third-party evaluators are other evaluation instruments that play a prominent role in teacher evaluation (Ross & Walsh, 2019). These practices originate from the Danielson Framework for Teaching (FFT;

2007), a commonly used framework for teacher evaluation used for teacher observations in the United States (Cleaver et al., 2018). Finally, with respect to the frequency, 22 of the 50 states in the USA require that all teachers be evaluated annually (Ross & Walsh, 2019).

In the case of Finland, teacher evaluation is characterized by a high level of teachers' autonomy. The purpose of the evaluation is a continuous improvement framed within the scheme of decentralization and trust in the schools and teachers' abilities, assuming that more trust will lead to improved teacher quality. Thus, Finland teacher evaluation based on professional development and teacher empowerment rather than a systematic tool of evaluation (Tarhan et al., 2019). The results of the evaluation are not used for accountability purposes and have little influence on their contractual status, however teachers are accountable to the community for the academic progress of all students. Accountability to the community operates through frequent meetings between teacher and parents and teacher committees to monitor all aspects of school life (Sahlberg, 2011).

Finnish Municipalities, which are responsible for running schools, are also in charge of developing the framework for the teacher evaluation of teachers who they employ. The framework is aligned with the requirements and guidelines put forward by the Ministry of Education. Therefore, Finland does not have a national framework for teacher evaluation (Tarhan et al., 2019).

Teachers in Finland are evaluated on their progress during a period based on the individual development plan that they prepare for themselves. Teacher evaluation is a consultative and formative process that usually takes place during a conversation between the teachers and their school principal or within a group of colleagues who teach the same

subject and grade. During the discussion, the teacher evaluates the fulfillment of the personal objectives previously established (Isoré, 2009).

For England, the evaluation system was originally designed with a summative purpose (Isoré, 2009). Teacher evaluation was based on teachers' performance on three different sets of standards according to the different stages of a teacher's career and the results were associated with a higher career level and economic incentive (Goepel, 2012). However, there were concerns about the potential problematic impacts of the process. Thus, the system made a change, incorporating an increased formative approach, that focuses on teacher professional development needs (Isoré, 2009), and that removed the element of progression (Cortez-Ochoa et al., 2018).

A national policy sets out the framework of a consistent teacher assessment of their overall performance. The new framework from 2012 is based on a single set of standards applied to teachers at any stage of their career (Goepel, 2012). The framework is made up of nine standards, eight related to teaching: 1) Set high expectations which inspire, motivate and challenge pupils; 2) Promote good progress and outcomes by pupils; 3) Demonstrate good subject and curriculum knowledge; 4) Plan and teach well-structured lessons; 5) Adapt teaching to respond to the strengths and needs of all pupils; 6) Make accurate and productive use of assessment; 7) Manage behavior effectively to ensure a good and safe learning environment; 8) Fulfill wider professional responsibilities, and one related to professional conduct: A teacher is expected to demonstrate consistently high standards of personal and professional conduct (Cortez-Ochoa et al., 2018).

Although the teacher evaluation system in England is regulated by basic principles at the country level, the responsibility for the system relies on collegiate bodies at the

school level. The national legislation in England gives flexibility to principals to modify the evaluation, according to what they consider most convenient for their teachers (Cortez-Ochoa et al., 2018). Thus, the evaluation is based on the standards but modified to the context of the school's plan for teacher improvement (Department for Education, 2019).

The evaluation process must be annual as a supportive and developmental process, designed to ensure that all teachers fully develop their skills and have access to the support they need to carry out their role effectively. Objectives for each teacher are specific, measurable, achievable, realistic, time-bound, and appropriate to the teacher's role and level of experience (Department for Education, 2019). The instruments used for the evaluation include observation and meeting with the principal, and in that meeting the teacher defines work objectives (Cruz-Aguayo et al., 2020).

In the case of Singapore, in 2001 the Ministry of Education replaced their evaluation system with a more comprehensive approach called the Enhanced Performance Management System (EPMS), which aims to help teachers better their performance. The system focuses on competencies related specifically to underlying traits and habits - patterns of thinking, feeling, acting, or speaking- that cause a person to be successful in a specific job. This competency model used in Singapore was based on David McClelland's research, who used a structured interview technique called the Behavior Event Interview (BEI) to elicit detailed stories that reveal how high performers differ from lower performing job holders. The high performers are used to develop a scale of increasingly effective behaviors associated with that competency (Steiner, 2010).

The teacher competency model developed in Singapore includes: a broad definition of the competencies that distinguish a high performance, rating scale of increasingly more

effective levels of behavior within the competency, and competency level targets for each level. This increasing level of competence enables teachers to perform better in the key results areas identified as critical to effective teaching in Singapore (Steiner, 2010).

The competency model in Singapore contains one core competency: Nurturing the Whole Child, and four competency clusters: Cultivating Knowledge, Winning Hearts and Minds, Working with Others, and Knowing Self and Others. Each cluster has between two to four competencies, that are broken down further into progressive levels that are used as rating scales. Each level includes a description of the specific behavior that the teacher has to demonstrate at a particular mastery level (Steiner, 2010).

Based on the competency framework, all teachers in Singapore develop annual performance goals at the beginning of the school year, using a standardized evaluation form. The form includes: goals that include competency targets; competencies; professional development plans for the next year; and feedback, that includes reviews and comments by the teacher and supervisor regarding work performance and competencies. When the teachers complete a draft of their evaluation form, they meet with a school supervisor officer at their school, who reviews the work making sure that it aligns with departmental, school, and national goals. Supervisors meet the teacher for midyear and final reviews, receiving constructive criticism about their goal achievement. At the end of the year, teachers meet with their supervisors to discuss whether they have met the goals (Steiner, 2010).

In the case of Latin American countries, teacher evaluation has generally not been a priority. Teacher evaluation is a topic that is debated between educational authorities and teacher unions. Its implementation is mediated by negotiations that do not always

encompass technical aspects of teachers' good performance based on the research related to teacher quality (Vaillant, 2008). Additionally, in a few countries of Latin America, evaluation has been accompanied by research on the process and the impact of the implementation, important information that could be used to improve the evaluation instruments used (Cruz-Aguayo et al., 2020). Therefore, evaluating teacher quality is a complex challenge, and few cases of national teacher performance evaluation can be found in Latin America (Vaillant, 2008). However, in the region there also some examples of teacher evaluation systems that can be considered "second generation", characterized by having a multiplicity of instruments and evaluators with the objective of understanding in detail all aspects of teaching processes, and ensuring validity and reliability of the system (Cruz-Aguayo et al., 2020). Between those systems, Chile and Colombia are mentioned very often in the literature, joined recently by Peru (Vaillant, 2008).

Since 2002, Colombia has implemented a new regulation that introduced permanent evaluation practices with the aim to ensure a continued satisfactory permanence by teachers, as well as providing incentives to improve over time. In order to achieve both aims, Colombia introduces two types of evolutions: 1) A yearly assessment for all teachers that is evaluated by the school principal, and reported to the local education authority, in which the principal comments on the teacher's performance following standardized criteria. Two consecutive years of negative evaluations lead to discontinuation of the employment as a teacher; 2) Competency evaluation for career upgrades conditional on passing public examinations that evaluate teachers' subject knowledge and teaching skills (Zelda & Sánchez, 2017).

The competency evaluation model contains four clear and pertinent criteria: the context of the educational and pedagogical practice of the teacher, reflection and planning of the educational and pedagogical practice, pedagogical praxis, and environment in the classroom. Through the competency model, the evaluation tries to identify the strengths and aspects to improve each teacher. The aspects that are to be improved are addressed through professional development that responds to the needs of teachers carried out by faculties of Education. Additionally, the new competency evaluation aims to create a system that eventually contributes to in-service training for teachers, taking into consideration a holistic evaluation of teacher practices and making sure not reduce that to a test as the unique evaluation source (Figueroa et al., 2018).

The Colombian competency model evaluation includes four evaluation instruments: a video of class that presents the teacher practices, the average of the teacher evaluation made up by the principal for the last two years, a student survey (from fourth grade onwards), and teacher self-evaluation. All the evaluation instruments and the correction rubrics are based on four criteria previously mentioned. The instrument construction process was carried out through joint work between the Colombian Institute for the Promotion of Higher Education (ICFES) and the Teaching Career project team. In addition, for its final review, it had the support of the MIDE center of the Pontifical Catholic University of Chile (MIDE UC; Figueroa et al., 2018).

In Peru, Teaching Performance Assessment is a formative and mandatory evaluation that seeks to transform and improve teacher practices in their classrooms and schools for the benefit of students served by the public education system. The Peruvian National Ministry of Education sets the evaluation criteria and instruments that will be

applied to assess teacher performance, so that an environment for professional reflection on their pedagogical practice is fostered. Likewise, the teachers evaluated receive specific feedback on their performance, which allows them to recognize their strengths and opportunities for improvement (Ministerio de Educación Perú, 2020).

Teacher evaluation in Peru is based on performance standards that were discussed and approved by various stakeholders. The standards are embodied in the Good Teaching Framework that was approved in 2012. This framework is made up of four domains: preparation for student learning, teaching for student learning, participation in school management in coordination with the community, and development of professionalism and teacher identity (Cruz-Aguayo et al., 2020).

Three types of instruments are used in Peru to evaluate teachers: classroom observation, student or parent surveys, and specific instruments to evaluate space management and responsibility. Teachers are evaluated every three years, and those teachers who fail the evaluation must participate in a professional development program led by the Ministry of Education (Cruz-Aguayo et al., 2020).

## Conceptualization of Validity

An effective teacher evaluation system must have technical validity, which means that it is able to distinguish between high-, average-, and low-performing teachers in a robust and consistent manner across different evaluators and over time. Moreover, the instruments used in the system should capture elements of teachers' skills and practice that are meaningfully linked to teachers' ability to help students learn and other important system goals. Systematic efforts to validate teacher evaluation systems are particularly important, especially with high-stakes evaluations (Bruns & Luque, 2014)

The *Standards for Educational and Psychological Testing* define validity as "the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests. Validity is, therefore, the most fundamental consideration in developing tests and evaluating tests" (AERA et al., 2014, p. 11). In order to evaluate a proposed interpretation of test scores for a particular use, we can use various sources of evidence. These different sources shed light on different aspects of validity, however they do not represent distinct types of validity (AERA et al., 2014).

The concept of validity emerged in the first half of the 20$^{th}$century, and it was defined as the extent to which any measuring instrument measures what it is intended to measure (Carmines & Zeller, 1979). However, over time, this concept has been modified. When this concept was first used, the validity of a psychological or educational test was evaluated by a diversity of procedures. The variance of these procedures depended on the test purpose, the theoretical orientation, and the availability of the data. Therefore, different researchers used a variety of names for the validity they reported, fluctuating between: face validity, validity by definition, intrinsic validity, logical validity, empirical validity, factorial validity, among others (Anastasi, 1986).

In 1954, the American Psychological Association, in an effort to introduce some order to these multiple definitions, published the *Technical Recommendations for Psychological Tests and Diagnostic Techniques,* which classified validity into content, predictive, concurrent, and construct validity. In future editions of this document, which eventually made up the *Standards*, predictive and concurrent validity were absorbed by criterion-related validity. With this change, the three most commonly defined types of validity were established and have survived until the present (Anastasi, 1986):

1. **Content validity**, a qualitative type of validity in which the concept domain is clearly defined and the measures fully represent the domain (Bollen, 1989). This type of validity concerns items sampling adequacy, and the extent to which a specific set of items reflects a content domain. In theory, a scale has content validity when its items are a random selection from the array of items. However, for some constructs, that could be more easily accomplished since the universe of items would be clearly defined. But, in the case of constructs such as beliefs, there is not a list of the relevant universe of items. A method used to maximize items' appropriateness, even in the case of the construct that is most difficult to figure out the universe of items, is the items review for the relevance of the domain by experts (DeVellis, 2017).

Content validity is closely linked to the definition of the construct being evaluated. A scale's content should reflect the conceptual definition applicable to the scale. It is imperative that item content capture the aspects of the phenomenon that are spelled out in its conceptual definition and no other aspects that might be related but outside the interpretation for a particular instrument (DeVellis, 2017).

2. **Criterion-related validity,** refers to the link between a measurement and a criterion variable. In order to assess criterion validity, we usually have a standard variable or "gold standard" that we compare to our measurement (DeVellis, 2017). If our criterion variables exist at the same time as the measure, it is defined as current validity. But, if the criterion occurs in the future, this is predictive validity (Bollen, 1989). However, the most important aspect of criterion-related validity is not the time relationship between the measure and the criterion variable, rather the strength of the empirical relationship between the two variables (DeVellis, 2017) .

3. **Construct validity**, refers to the theoretical relationship between one variable and another variable (DeVellis, 2017). Construct validity is concerned with the extent to which a particular measure relates to other measures consistent with theoretical constructs that are being measured (Carmines & Zeller, 1979). In contrast with criterion validity that is often assessed by the correlation between the measure and the criterion, construct validity can be assessed only indirectly because the relevant comparison is to a latent variable rather than an observed variable (DeVellis, 2017).

Although the tripartite categorization of validity initially helped to clarify validation procedures, it had some adverse effects on testing practices. At times, this categorization oversimplified grouping of data-gathering procedures, leading to a superficial understanding of test measures. In an attempt to overcome this distortion of the role of validity, the 1985 edition of the *Standards* eliminated some of the apparent rigidities of the earlier editions, opting for a more comprehensive approach to validation procedures (Anastasi, 1986).

Therefore, since validity was first defined, several changes in its focus and emphasis have taken place (Angoff, 1988). Some authors have given validity concepts a broad definition. For instance, Messick (1995) and Kane (2013), among others, pointed out that the concept of validity is related to whether a specific test score interpretation or use is valid. Validation is an ongoing process and is judged in terms of degrees and never in absolute terms (Messick, 1995). Another important change in validity conception was the idea of not validating the test, rather the data interpretation arising from a specified procedure (Cronbach, 1988). Therefore, the measuring instrument is not validated itself, rather the measuring instrument it is measure with respect to the purpose of its use.

Taking into consideration this definition of validity, there are several different types of validity that take a somewhat different approach in assessing the extent to which a measure measures what it purports to (Carmines & Zeller, 1979).

The evolution of the validity conceptualization was also reflected in the *Standards* definition of validity in 1985. Since that edition, the *Standards* have incorporated the importance of the evidence that supports test result inferences in their test validation definition. For the *Standards,* evidence of validity can include different sources that might be used in validity evaluation of a particular aspect of a test score. Those sources may shed light on different aspects of validity, but they do not represent different types of validity (AERA et al., 2014). Therefore, validity is a unitary concept based on various kinds of evidence (Miller et al., 2009). It is conceptualized by the *Standards* as the degree to which all of the accumulated evidence supports the intended interpretation of the test purposes used. As we see, this framework does not follow the historical nomenclature of validity already presented (AERA et al., 2014).

The *Standards* define five types of evidence and each one is not required in all settings, rather we use different evidence types as needed. Therefore, not all instruments require the same type of evidence, and depending on the intended use of the instrument, they may require more validity evidence than others (AERA et al., 2014).

The five types of evidence described by the *Standards* are:

1) Evidence based on test content, obtained from the analysis of the relationship between the test content and the construct it is intended to measure. This type of evidence can include a logical or empirical analysis of the adequacy with which the test content

represents the content domain, and experts in the field consider the relationship between parts of the test and the construct that is measured.

2) Evidence based on the response process, related to the fit between the construct and the way in which test takers engage with the evaluation. In general, the analysis source comes from individual responses. Asking test takers about their strategies used to respond to a particular item can yield evidence that enriches the construct definition.

3) Evidence based on internal structure, indicates the degree to which the relationship between test items and test components conform to the construct, on which the test score interpretations are based.

It is possible to describe three basic aspects of the internal structure of an instrument: dimensionality, measurement invariance, and reliability. Dimensionality explores whether or not the inter-relationship among the items supports the intended test score used to make inferences. For instance, a test that reports one composite score should be predominantly unidimensional. Measurement invariance provides evidence that item components are comparable across different specific groups, such as gender or race. Lastly, reliability refers to evidence that the test scores are consistent across repeated measurements (Rios & Wells, 2014).

Therefore, assessing dimensionality is one of the main aspects of internal structure validity (Rios & Wells, 2014). The first step in the achievement test development process is to define whether it is unidimensional or multidimensional (Haladyna & Kramer, 2004). For instance, when empirical test multidimensionality coincides with the hypothesized structure of the test, it proves the internal structure validity of the total score (Tate, 2002).

The *Standards* (AERA et al., 2014) indicates that in order to be able to claim that a test is unidimensional, such a claim must be supported by multivariate statistical analysis, such as factor analysis. The analysis should show that the score variability attributable to one major dimension is much greater than that of other identified dimensions. Given its widespread use, factor analysis has been used as the statistical method to assess test dimensionality in a particular set of data (Brown, 2015). However, there are several other analytical methods available for analyzing test dimensionality (Rios & Wells, 2014).

4) Evidence based on the connection between variables, connects the construct to some other external variables on the test. This type of evidence may include measurements of a particular criterion that the test is expected to predict: the same or similar construct (convergent evidence), or test measurement related to a different construct (discriminant evidence).

5) Evidence for validity and consequences of testing, focuses on the interpretation of test scores for intended or unintended uses by the test developer (AERA et al., 2014).

Despite the conceptualization of test validity being more related to the tests themselves, the concept has evolved towards a conception of validity related to the purpose of the test. Therefore, the validity of a test is related to different types of evidence that can vary with the purpose of the test (Anastasi, 1986). In educational measurement settings, the term validity has been used in this second sense, which means that its validity is not characteristic of the test itself rather it depends on the particular use of the test. Different uses may entail different inferences from the test results (Koretz, 2008).

Taking into consideration this conceptualization of validity, the present research seeks to contribute to the body of research on the evidence of the validity of the portfolio used for the Chilean teacher evaluation system. As previously indicated, the evaluation in Chile has at least two clearly defined purposes: providing formative data on individual teachers to improve their practice, and providing summative data to support individual teacher reward and sanctions (Taut et al., 2012). In order to support both aims, the test scores from the portfolio are interpreted as an indicator of a teacher's overall quality of teachers' competences. Concurrently, the portfolio is a guide for identifying areas of teacher improvement. Therefore, evidence of validity for the portfolio in the present research focuses on correctly distinguishing overall teacher quality for teachers from different contexts, with results based on different components of the portfolio (taking into consideration Collaborative Work or not), and from a specific portfolio weighted score. Finally, evidence of the portfolio validity also focuses on a valid way to identify teacher strengths and weaknesses.

## Evidence of Validity of Teacher Evaluation Systems

A comprehensive validation effort regarding large scale teacher assessment systems has not been documented extensively in the literature. One exception was found by the National Board for Professional Teaching Standards (NBPTS) certification of teaching excellence. NBPTS performed assessments to certify accomplished teachers in the United States. Thus, their psychometric evaluation is critical to the program's effectiveness (Hakel et al., 2008). With respect to validity, there have been three types of studies to gather content-based validity evidence. The first study investigated the processes used to develop content standards. In this study, researchers examined the

extent to which the development of standards had a scientific basis (Hattie, 2008). The second study evaluated the congruence between the assessment and its content domain. The study relied on an expert panel to judge the appropriateness of the domain, defined by the assessment content standards, and whether the scoring represents the intended content domain (Crocker, 1997). The third study focused on the scoring rubrics. In this study, panelists reviewed a series of pairs of exercise responses. After that, they were asked to review the content standards for the assessment and to make judgments about which of each pair of responses should receive the higher consistency score with respect to the standards (Jaeger, 1998).

The board for the NBPTS assessment also collected construct-based validity evidence. The most important study involved classroom observations, evaluating the extent to which board-certified teachers demonstrate the knowledge, skills, dispositions, and judgments, both in their practices as well as on the assessment. The researchers compared the performance of two groups: board-certified teachers and unsuccessful applicants using 15 key dimensions of teaching expertise. They found that board-certified teachers scored higher on all of these dimensions (Hakel et al., 2008).

Another evidence of validity of the teacher evaluation system has been reported to several implementations of Danielson's Framework in the United States. Validity studies of the framework have primarily focused on *criterion-related validity* evidence, in which the researchers have studied the relationship between teachers' evaluation ratings and teachers' effects on student learning, as represented by classroom-level value-added estimates of teacher productivity. This type of evidence is based on the idea that if there

is an external standard of performance (the criterion), then ratings should correlate with or predict measures of the standard (Milanowski, 2011).

While *criterion-related validity* evidence is not the only type of validity evidence that matters, it has been the most commonly sought out. Results from this study indicate that teachers in the top value-added quartile consistently received higher ratings on all of the standards, and a one point increase in the average of evaluation ratings on the standards is associated with a one-sixth standard deviation increase in math achievement and a one-fifth standard deviation increase in reading achievement, controlling for other variables (Milanowski, 2011).

## Evidence of Validity of Chilean Teacher Evaluation System

A comprehensive validation plan for the Chilean National Teacher Evaluation System was developed from 2005 to 2012. The research was mainly conducted by researchers from MIDE center of Pontifical Catholic University of Chile (MIDE UC), which was also responsible for developing and implementing the Chilean National Teacher Evaluation System, developed in 2005 as a long-term research agenda to gather evidence for the validity of the system (Taut et al., 2012). The validity plan included a variety of methods and sources of evidence, and given the limited budget, they prioritized the most relevant studies for the validity of the program (Taut et al., 2011).

The studies started with the *Standards* (AERA et al., 2014) as a starting point to organize the validation work around the types of evidence delineated there. Therefore, questions related to different types of evidence of validity, such as content validity, internal structure, the relationship between variables, and consequential validity, were the main topics from which the researchers structured their work (Taut et al., 2012).

For content validity of the Chilean National Teacher Evaluation System, the researchers studied alignment of the system with the standards for good teaching described in the Good Teaching Framework (GTF). This study found that the portfolio covered a large majority of indicators related to Domain B and C, and partially covered those related to Domain A and D. Nevertheless, other instruments used in the evaluation (peer interview, third party report, and self-evaluation) assessed indicators related to Domain D. For Domain A, they found a limited coverage of standards related to subject-specific pedagogy and content knowledge, which was also not addressed by the other evaluation instruments used in the evaluation system (Taut et al., 2012).

For validity evidence based on relationships with other variables, the researchers conducted several studies on the relationship between the Chilean National Teacher Evaluation System results and other variables measuring similar constructs. The first study compared the pedagogical practices of teachers who have high scores with those who have low scores on the evaluation (Santelices & Taut, 2011). A second set of studies looked on evidence for the relationship between teacher results on the Chilean National Teacher Evaluation System with the student achievement according to the Chilean Education Quality Measurement System (Sistema de Medición de Calidad de la Educación, SIMCE[9]). Finally, a third study explored the relationship between teachers' results and their scores on another voluntary teacher evaluation program (AEP[10]), which

[9] SIMCE is a mandatory national standardized assessment of student performance. It measures student performance in language, mathematics, and science (in grades 2, 4, 6, 8 and 10) and in English (grade 11). The results are widely publicized and are used to allocate resources and rewards to schools and teachers, to guide educational policy, and to provide information to parents.

[10] AEP was a complementary teacher evaluation process that recognized professional merit using subject matter and pedagogical tests, as well as a portfolio, based on the Good Teaching Framework.

also involved a portfolio based on the Good Teaching Framework (GTF; Taut et al., 2012).

The general results for all of those studies backed up the evidence that supports the validity of the Chilean National Teacher Evaluation System scores (Taut et al., 2012). First, the teacher evaluation accounts for real differences between teachers with high and those with lower performance. The study reported substantive differences between the pedagogical practice's performance between teachers with unsatisfactory and outstanding results. Second, the studies showed positive correlations between teacher performance and student achievement. Finally, the results comparing teachers evaluated by two similar programs (Teaching Evaluation and AEP), both based on GTF and in the use of portfolios, provide additional positive evidence regarding convergent validity (Taut et al., 2011).

With respect to the validity evidence related to the consequences of the teacher evaluation, the researchers used mixed methodologies to examine empirically the possible effects of the Chilean National Teacher Evaluation System on the teachers evaluated. They evaluated the consequences of the teacher in terms of the teacher's participation in professional development or their participation in the voluntary program Variable Individual Performance Allowance (AVDI *in Spanish*). The researchers also conducted interviews and focus groups to find the expected and unexpected consequences of the evaluation for the local school board (municipalities), principals, and teachers (Taut et al., 2011).

The findings indicated that at the municipal level, the results are used for decision making related to teachers' professional development and for local recognition of good

teachers. At the school level, positive effects were observed in terms of the promotion of collaboration between the teachers evaluated. Finally, at the individual level, the effects were mixed. There was a general perception of negative emotional reactions for the teachers evaluated due to the work overload that the complete process implied. However, teachers also recognized important benefits of the evaluation process, especially in the development of the portfolio, in terms of reviewing and updating their practices (Taut et al., 2011).

For validity evidence based on internal structure of the Chilean National Teacher Evaluation System, from 2005 to 2010, MIDE UC researchers used exploratory and confirmatory factor analysis to study the structure of the evaluation instruments, with particular emphasis on the study of the portfolio structure (Valencia & Taut, 2008). The portfolio used for these studies had a different structure from the portfolio that is currently used. The structure of the portfolio contained 24 indicators, grouped into eight dimensions, each one associated with a product from Module 1 or Module 2 (five for Module 1 and three for Module 2). Table 2.1 presents the eight dimensions and their association with each module evaluated.

Table 2.1.
*Dimensions evaluated by the Portfolio before 2016*

| Dimension | Domain |
|---|---|
| A: Organization of the elements of the learning unit<br>B: Analysis of class activities | Module 1: Pedagogical Unit |
| C: Quality of the assessment of the learning unit<br>D: Reflection on assessment results | Module 1: Pedagogical Unit Assessment |
| E: Reflection on pedagogical practices | Module 1: Reflection |
| F: Classroom environment<br>G: Structure of the class<br>H: Pedagogical Interaction | Module 2: Video recorded class |

*Note.* Translated and adapted from (Flotts & Abarzua, 2011)

The eight dimensions presented above had the same relative weight in the calculation of the teacher's portfolio final score. The score for each dimension was reported to each teacher in their final report. The results were presented in terms of a performance category (unsatisfactory, basic, competent, or outstanding), which was similar in which the way to the other evaluation instruments result were reported to the teachers (Flotts & Abarzua, 2011).

Therefore, taking into consideration the structure of the portfolio presented in table 2.1., the researchers answered questions regarding the number and nature of latent variables that might explain shared variances of a matrix of correlations of portfolio indicators, through exploratory factor analysis. They applied the Principal Axis Factoring and Maximum Likelihood estimation methods, as well as various factor retention rules

(Kaiser-Gutmann, screen test, interpretability). The rotation method that the researchers used was Oblimin (Taut et al., 2012).

Results from the annual exploratory factor analysis from the years 2005 to 2010 varied somewhat over the six years analyzed. In general, the results identified either five or six factors for the entire portfolio, including the pedagogical material and the video recorded class. These factors explain between 30% to 39% of the variance in scores. Usually, three of the factors were associated with the pedagogical material part of the portfolio, and the other factors were associated with the video recorded class. The factors associated with the video recorded class varied from year to year, however for some years they neatly re-created the underlying theoretical dimensions of the portfolio. On the other hand, the factor associated with the pedagogical material part was more stable over time. Overall, exploratory factor analysis identified the factor structure that resembled the portfolio modules more than the dimensions (Taut et al., 2012).

With the results from the 2010 Chilean National Teacher Evaluation System portfolio (10,350 observations), the researchers used confirmatory factor analysis looking for evidence regarding the pre-established underlying portfolio structure. For this analysis, they used M*plus* 5.21 to apply robust weighted least squares estimation method and a tetrachoric correlation input matrix., taking into consideration that data were ordered nominal. In order to evaluate the model fit, the researchers considered indexes such as: chi-square test of model fit, comparative fit index (CFI), Tucker-Lewis fit index (TLI), and root mean square error of approximation (RMSEA), with the appropriate cutoff values proposed by Brown (2015). The results from portfolio confirmatory factor

analysis indicated that the theoretical structure of the eight dimensions fits the data well, according to CFA indexes (CFI, TLI, and RMSEA; Taut et al., 2012).

The results for validity evidence based on the internal structure of the Chilean National Teacher Evaluation System portfolio partially validate the structure of that evaluation instrument, providing suggestions for improving future scoring and reporting (Taut et al., 2012). The findings from the validity evidence based on the portfolio internal structure were, in part, taken into account by the portfolio structure changes made to the application of the evaluation from 2016 onwards (Sun, 2018).

## Limitation from previous research

The agenda of validation carried out by the MIDE UC researchers was shaped in part by the resource constraints that they confronted. Therefore, they had to prioritize those studies that seemed most crucial to investigating for the system: first, the proposed interpretations and uses of the Chilean Teacher Evaluation System, which was the basis of their validation work that came out of their empirical work analyzing policy documents and interviewing relevant stakeholders. They focused on the most important purposes and uses. Second, they decided to focus on the overall assessment score because it has direct consequences for individual teachers, schools, and municipalities, and the portfolio instrument because it has the most weight in the determination of the final score (Taut et al., 2012).

However, the previous research from the validation of the Chilean Teacher Evaluation System was interrupted in 2012. Since 2012, important changes have occurred in the process of teacher evaluation in Chile, without any further formal validity processes. Between the recommendations presented by Manzi & Jiménez ( 2017), in the

context of a presentation of the evidence that supports the validity of the teacher evaluation in Chile in MIDE UC, they indicate the need to design a validation program for the instruments. This need is in the context of the inclusion of the new Teacher Education and Professional Development as part of the Teacher Evaluation System and also given the change in the use of the instruments and the increase in their consequences.

The portfolio has been changed since 2016 with the inclusion of Module 3. Also, there are standardized tests of disciplinary and pedagogical knowledge that have been taken by all the teachers in their career progression (those tests were voluntary before). Finally, the inclusion in the evaluation of all the teachers that work in charter schools should be evaluated (Manzi & Jiménez, 2017). All of these changes mentioned above present challenges regarding the validation process of the Teaching Evaluation in Chile. Although not all will be addressed in the present research, this may be a first step in this new validation agenda.

# Chapter 3: Methodology

For the present research, I am using the data results from the 2017 teacher evaluation carried out by the Chilean National Teacher Evaluation System. Teacher evaluation results from 2004 to 2017 are available to the public from the Research Center of the National Education Ministry. However, the information about indicators and domains from the 2017 portfolio evaluation is not available to the public on the Research Center webpage. In order to access this data, I requested and received the data using the Chilean Transparency Law, and as a result, I have obtained the necessary information for this dissertation.

The focus of the current chapter includes: a description of the sample used for the analysis, a description of the measurement instrument, and the analytic strategy to address the aims.

## Data Source

This dissertation was conducted using the dataset from the results of the 2017 Chilean National Teacher Evaluation. This dataset included results from 24,251 teachers who worked in the Chilean municipal schools throughout the 15 regions of the country. This teacher cohort corresponded to 23.34% of the population of teachers who worked in the municipal schools in the year 2017. The teachers evaluated in 2017 taught in the areas of early childhood, elementary school, middle school, high school, special education, adult education, and technical education.

The main goal of this dissertation is to contribute to the body of research on the evidence of the validity of the portfolio used by the Chilean Teacher Evaluation System. The structure of the portfolio is the same for all of the teachers evaluated, regardless of the grade level or subject taught, with the exception of the technical education teachers. For those teachers, the indicators that make up the portfolio are slightly different in order to align the teacher evaluation with the technical teachers' work. Given this difference, for the present research, technical education teachers were removed from the sample for the analysis (*n*= 2,269; 9.4% of the total sample). Therefore, the total analytic sample used for the analysis was 21,982 teachers.

Table 3.1 presents the descriptive statistics of the teachers sample. Approximately 74% of the teachers evaluated in 2017 were female. The average age of the teachers was 40 years old (*SD* = 11.21; range 22-79). They were primarily Chilean (99%), and did not identify with any ethnic group[11] (96%). With respect to the level that they taught, 7% were early childhood teachers; 20% elementary teachers; 32% middle school teachers; 19% high school teachers; 20% special education teachers; and 2% adult education teachers. In terms of school location, 78% of the teachers worked in urban schools, and 22% in rural schools. Teachers who self-reported having a professional degree in education were 64%, while 36% were unreported. The breakdown of the teachers who reported having a professional degree is: 5% early childhood; 29% elementary school; 19% high school; and 11% special education. With respect to the subject matter that the

[11] The State of Chile recognizes that ethnic refers to the people of Chile and are the descendants of the human groups that have existed in the national territory since pre-Columbus time, that preserves their own cultural manifestation (Law 19,253, 2017). According to the latest available data, the ethnic population represents 4.6% of the total Chilean population (INE, Census 2002).

teacher taught, 7% taught Preschool; 20% Elementary (1st to 4th grade, all subject areas); 9% Language Arts; 9% Math; 6% History; 6% Science; 6% Foreign Language; 7% Physical Education; 2% Music; 2% Art; 1% Technology; 3% Religion; 1% Philosophy; 20% Special Education; and 2% Elementary and High School Adult Education.

Table 3.1.
*Demographic Information about the Chilean National Teacher Evaluation 2017 (N=21,982).*

| | Frequency | Percent |
|---|---|---|
| Teacher's Gender | | |
| Male | 5,680 | 25.84 |
| Female | 16,302 | 74.16 |
| Citizenship | | |
| Chilean | 21,831 | 99.32 |
| Foreigner | 73 | 0.33 |
| Ethnic Identification | | |
| Yes | 793 | 3.61 |
| No | 21,120 | 96.08 |
| School Location | | |
| Urban | 17,040 | 77.52 |
| Rural | 4,942 | 22.48 |
| Level Taught | | |
| Early Childhood | 1,497 | 6.81 |
| Elementary School | 4,385 | 19.95 |
| Middle School | 7,019 | 31.93 |
| High School | 4,309 | 19.60 |
| Special Education | 4,325 | 19.68 |
| Adult Education | 447 | 2.03 |

| | Frequency | Percent |
|---|---|---|
| Subject Matter Taught | | |
| Early Childhood | 1,497 | 6.81 |
| Elementary (1st to 4th) | 4,385 | 19.95 |
| Math | 1,884 | 8.57 |
| Language | 1,888 | 8.59 |
| Science | 1,339 | 6.09 |
| Physical Education | 1,508 | 6.86 |
| Music | 514 | 2.34 |
| Art | 476 | 2.17 |
| Technology | 255 | 1.16 |
| Foreign Language | 1,309 | 5.95 |
| Religion | 719 | 3.27 |
| Philosophy | 150 | 0.68 |
| Special Education | 4,325 | 19.68 |
| Elementary and High School Adult Education | 447 | 2.03 |
| Certification (Self-Reported) | | |
| Early Childhood | 1,213 | 5.51 |
| Elementary School | 6,364 | 28.95 |
| High School | 4,094 | 18.62 |
| Special Education | 2,462 | 11.20 |
| Unreported | 7.849 | 35.71 |

**Missing Data**. As previously described, the dataset with the results of the 2017 Chilean National Teacher Evaluation is composed of 24,251 observations. Taking into consideration the removal of technical education teachers due to the differences in the portfolio indicators, the analytic sample decreases to 21,982 teachers (9.4%).

In order to explore how different the sample of technical teachers that were removed from the analysis with respect to the analytic sample was, I conducted a series of tests between the technical education teachers and the sample used for the present

research (i.e., 21,982 vs. 2,269), using the demographic variables presented in Table 3.1. The results indicate significant differences across the analytic sample with the technical education teachers for gender, ethnic identification, school location, and age. For gender, the analytic sample included a smaller proportion of male teachers in comparison with the technical education teacher sample (25.84% vs. 60.69%, $\chi^2$ (1, $N$= 24,251) = 1,200, $p$ < 0.01). The proportion of teachers that identified as belonging to any Chilean ethnic group was slightly higher for the analytic sample (3.62% vs. 2.43%, $\chi^2$ (1, $N$= 24,176) = 8.56, $p$ = 0.003). For school location, the proportion of teachers working in rural schools was higher in the analytic sample compared to the technical education teacher sample (22.48% vs. 5.07%, $\chi^2$ (1, $N$= 24,251) = 377.88, $p$ < 0.01). Finally, with respect to the teacher's average age, teachers from the analytical sample were significantly younger than the teachers from the technical education sample ($M$ = 39.66, $SD$ = 0.08; $M$ = 43.56, $SD$ = 0.23, respectively; $t$(24,249) = -15.79, $p$ < 0.01).

These indicators are very similar to the statistics from the Ministry of Education, in which they describe the teachers of technical educational schools. Unlike what happens in the other educational levels, the technical educational teachers are predominantly male (close to 60%), with an average age of 44.7 years old. One of the characteristics of technical teachers reported by the Ministry is that around half of them do not have a certification in pedagogy, rather a professional or technical certification in the subject they teach. That goes hand in hand with the fact that within the characteristics of the curriculum to be taught by these teachers is the deepening of the students' practical learning, which implies the need for teachers favored by teachers with practical experience beyond the pedagogical certification (Sevilla, 2011). Therefore, these analyses

help clarify that technical teachers present some specifically different characteristics when they are compared to the remaining teachers evaluated, which is consistent with the use of a differentiated portfolio indicator. This is also helpful for the justification for the removal of technical teachers in the present research.

Missing information for the analytic sample ($N$ = 21,982) was also analyzed. The results indicated that the percentage of missing data was 2.5% ($N$ = 550) for all indicator variables evaluated in the portfolio (Module 1, 2, and 3). For Module 1 indicators, the percentage of missing data ranged from 2.5% to 3.4%, the indicator 3.1 (*Analysis and use of assessment results*) being the indicator with higher missing information ($N$ = 835). With respect to Module 2, the missing data was the same in all of the nine indicators, corresponding to 559 teachers (2.5%). Finally, Module 3 indicators have the highest percentage of missing information, fluctuating between 18.33% and 19.03%. The reason for the higher percentage of missing data in Module 3 is because this module is optional and the results are only taken into consideration when it benefits the teacher. Therefore, many of the teachers do not turn in evidence to be evaluated for this module.

In order to assess the missing data, I first conducted Little's MCAR test to check for missing cases for each one of the modules being missing completely at random. The results of the p-value for Little's MCAR test were significant for each module, indicating that the data were not missing completely at random. Later, I conducted a series of tests between the missing and non-missing groups, using the demographic variables presented in Table 3.1, to test missing at random (MAR). When I compared the differences from the analytic sample and the missing cases for each one of the modules, I found a common pattern among the differences. Over the three sets of comparisons, the significant

differences were in gender, school location and teacher age. For gender, the analytic sample included a smaller proportion of females when compared with the missing cases for Module 1 and Module 2 (74.05% vs 78.55%, $\chi^2$ (1, $N$= 21,982) = 5.66, $p$ = 0.017; 74.06% vs. 78.18%, $\chi^2$ (1, $N$= 21,982) = 4.82, $p$ = 0.028, respectively). However, for Module 3, the proportion of female teachers was significantly larger in the analytic sample (76. 26% vs. 63.17%, $\chi^2$ (1, $N$= 21,982) = 310.44, $p < 0.001$). In the case of school location, the differences were similar for the 3 modules; the proportion of teachers working in rural schools was significantly larger for the analytical sample (22.64% vs. 16.36%, $\chi^2$ (1, $N$= 21,982) =12.12, $p < 0.001$; 22.63% vs. 16.64%, $\chi^2$ (1, $N$= 21,982) =11.24, $p$ = 0.001; 22.83% vs 20.92%, $\chi^2$ (1, $N$= 21,982) = 6.86, $p$ = 0.009, respectively). Finally, for the average teacher age, teachers from the analytical sample were significantly younger than the teachers with missing information in the 3 modules ($M$=39.43, $SD$=0.08; $M$=48.68, $SD$=0.38; $t$(21,980)= -19.27, $p < 0.001$; $M$=39.43, $SD$=0.08; $M$=48.68, S$D$=0.37; $t$(21,980)= -19.44, $p < 0.001$; $M$=39.24, $SD$=0.08; $M$=41.54, $SD$=0.19; $t$(21,980)=-11.80, $p < 0.001$, respectively).

The results indicate that missingness in all the indicators was significantly correlated to gender, school location, and teacher age. Therefore, missingness may be predicted to some degree, consistent with missing at random (MAR; Schafer & Graham, 2002). Taking into consideration these results, the low percentage of missing data at least for Module 1 and Module 2, and the problem of using multiple imputation for the analysis proposed by the present research, the decision for dealing with missing data was to do pairwise deletion. This method for dealing with missing data attempts to mitigate

the loss of data by eliminating cases on an analysis-by-analysis (Enders, 2010). Pairwise deletion is known to be less biased for the MCAR or MAR data, but if there are many missing observations, the analysis will be deficient (Kang, 2013). Therefore, it is important to consider that the caveat results from case deletion may be biased since the complete cases can be unrepresentative of the full population. This could be particularly problematic for the indicators of Module 3 because the percentage of missing is close to 20%. This possible bias problem was considered in the analysis and discussion.

## Instrument

The Chilean National Teacher Evaluation evaluates teachers from municipal schools every four years. They use four instruments to obtain a holistic view of the teacher: self-evaluation, peer evaluation, third-party assessment, and portfolio practice assessment. The focus of the present dissertation is on this last instrument, as it is the core tool of the teacher evaluation.

**Portfolio.** The portfolio is one of the four instruments used by the Chilean National Teacher Evaluation System to evaluate the teachers, and it is part of the new Teacher Professional Development System. This instrument allows for the collection of concrete evidence of teacher practices, which helps shed light on the actual performance of the teacher in the classroom.

The portfolio is grouped into three modules, all of which require teachers to submit different evidence for their evaluation. The first module evaluates planning, evaluation, and reflection of a teaching unit. The second module evaluates a 40-minute video recorded class. Lastly, the third module evaluates teachers' collaborative work.

*Module 1*. This module refers to the teachers' submission of a series of written documents of pedagogical materials. This module includes three dimensions: planning, assessment, and reflection, from which there are seven total indicators.

The first dimension of Module 1 contains two indicators: formation of learning objectives and the relationship between activities and objectives. In this module, the teacher has to include the planning of three classes within a specific unit. The task that the teacher is asked to present encompasses how they address curriculum objectives and content, and how their students approach learning and acquiring new skills and knowledge. The teacher must implement this pedagogical unit in the classroom. The unit must be based on one of two grade level content options from the Chilean National Educational Standards, which are reviewed and updated each year. Once teachers choose the unit, they plan the entire unit starting with the overarching objectives, eventually focusing on the objectives for each of the classes that comprise the unit. Teachers must submit the planning for three classes from this pedagogical unit, indicating the date, duration, activities carried out, resources used, among other aspects. Table 3.2 outlines the planning dimension indicators with the performance standards for a competent teacher (benchmark).

Table 3.2.
*Indicators and Performance Standards for the Planning Dimension*

| Dimensions | Indicators | Performance Standards for a Competent Teacher |
|---|---|---|
| Planning | 1.1. Formation of learning objectives. | The teacher demonstrates clear learning objectives by identifying both the skills and contents that the students must develop. |
| | 1.2. Relationship between activities and objectives | The teacher carries out activities that allow the students to achieve the learning objectives, and that covers both the skills and contents previously developed. |

*Note.* Translated and adapted from www.docentemas.cl

The second dimension for Module 1 is made up of three indicators: evaluation and rubrics used for correction, the relationship between assessment and objectives, and the analysis and use of assessment results. For this module, the teachers have to include the students' learning assessment from that pedagogical unit, accompanied by the rubric correction. The main requirement of this evaluation is that it measures what students have learned in that particular unit. If it is a written test, the teacher must send a copy of the test with the correct answers marked, or conversely, the criteria used to evaluate each response. If the teacher uses a different assessment, for example, playing a musical instrument or an oral presentation, they must describe the evaluation process and the instructions that were given to the students. In addition, the teacher must present the correction guide used to assess their students. Along with the assessment, teachers

present an analysis of the results of the evaluation. From this analysis, they reformulate and adapt teaching activities to improve learning. Thus, for this indicator, teachers should analyze the students' results in the assessment and propose educational activities according to these results. Table 3.3 presents the assessment dimension indicators with the performance standards for a competent teacher.

Table 3.3.
*Indicators and Performance Standards for the Assessment Dimension*

| Dimensions | Indicators | Performance Standards for a Competent Teacher |
|---|---|---|
| Assessment | 2.1. Evaluation and rubrics used for correction | The instructions, questions, or tasks included in the evaluation are clear. Additionally, the rubric correctly identifies the expected performance. |
| | 2.2. Relationship between assessment and objectives | The different evaluations are consistent with the learning objectives intended to be measured. |
| | 2.3. Analysis and use of assessment results | Based on the student assessment results, the teacher does an in-depth analysis of both learning outcomes and their causes. Additionally, the teacher uses this analysis to propose pedagogical strategies for improving student learning. |

*Note.* Translated and adapted from www.docentemas.cl

The third aspect of Module 1 is composed of two indicators: analysis based on students' characteristics and the use of error for learning. This module evaluates how the teacher incorporates their students' personality and needs into the unit planning. The teacher also has to foresee common difficulties that students could possibly face, which is a formative approach. In this task, the teacher must reflect on how they addressed any difficulty or relevant error that they observed. The reflection dimension indicators and the performance standards for a competent teacher are included in Table 3.4.

Table 3.4.
*Indicators and Performance Standards for the Reflection Dimension.*

| Dimensions | Indicators | Performance Standards for a Competent Teacher |
|---|---|---|
| Reflection | 3.1. Analysis based on students' characteristics. | The teacher demonstrates awareness of their students' characteristics and takes them into consideration when planning or teaching their class, looking out for teachable moments. |
| | 3.2. Use of error for learning | The teacher identifies a student's error that is relevant to their learning process and is able to fully understand why it occurred. Using this information, the teacher walks the student through targeted strategies so that they can understand their mistake and improve their performance. |

*Note.* Translated and adapted from www.docentemas.cl

*Module 2.* This module contains only one dimension, in which there are nine indicators: class environment, quality of the start of class, quality of the end of class, the contribution of the activities to the achievement of the class objectives, curricular emphasis on the subject, clear explanations, questions and activities, encouragement, and student feedback. For this part of the evaluation, the teacher must arrange a time with the school principal to have a 40-minute class video recorded. The teacher is expected to show their effectiveness in carrying out a class, demonstrating a clear start, middle, and end of a lesson. In addition to presenting the filmed class, the teacher must complete a brief description of the aspects of the class and attach any learning resources used, as these may not be clearly seen in the video. Table 3.5 presents the indicators for Module 2, with the performance standards for a competent teacher.

Table 3.5.
*Indicators and Performance Standards for Module 2: Video Recording of a Class*

| Indicators | Performance Standards for a Competent Teacher |
|---|---|
| 4.1. Class environment | The teacher uses effective strategies to get their students to do the activities in a respectful environment. The teacher shows interest in what their students do and say, generating an environment of trust for them to make mistakes, disagree, raise concerns, and so on. |
| 4.2. Quality of the start of class | At the beginning of class, the teacher motivates their students to engage with the material for that class, connecting previous learning to the objective for that day. |
| 4.3. Quality of the end of class | At the end of the class, the teacher carries out a closing activity that either summarizes, applies, or deepens the class content in order to consolidate the material learned that day. |

| | |
|---|---|
| 4.4. Contribution of the activities to the achievement of the class objectives | The teacher addresses all of the proposed objectives through the activities carried out in the class, taking advantage of all the time available to devote it to learning. |
| 4.5. Curricular emphasis on the subject. | The strategies the teacher implements are consistent with the subject matter taught. |
| 4.6. Clear explanations | The teacher's explanations connect with their students' previous experiences. For example, the teacher uses previous knowledge to explain and deepen a new concept. In the case of explaining a procedure or skill, the teacher tries to ensure that students understand the best way to do it. |
| 4.7. Questions and activities | The questions and activities presented to the students are challenging. They motivate the students to analyze, interpret, create, or apply them to the classwork, rather than repeating or paraphrasing information. With these learning skills, the teacher promotes the development of higher order thinking skills in their students. |
| 4.8. Encouragement | The teacher helps the students participate in an active and equitable way throughout the class, and encourages student interaction that promotes peer learning. For example, the teacher encourages pair work in which students help and explain their ideas to each other. |
| 4.9. Student feedback and use of assessment results | During the video recorded class, the teacher provides feedback to their students, allowing them to learn from their own work. This process encourages them to add to their responses, analyze the steps they took to arrive at the answer, or to identify the reason for their success or failure. |

*Note.* Translated and adapted from www.docentemas.cl

*Module 3.* This non-mandatory module is made up of one dimension with four indicators: collaborative work suitability, quality of professional dialogue, value of

collaborative work for professional development, and reflection on the impact of the collaborative work experience. With the creation of the Teacher Professional Development System in 2016, this new module of the portfolio was established. The idea was to enrich the portfolio by incorporating new aspects or dimensions, such as collaborative work, professional development, and other professional responsibilities that the teacher takes on in the school. Given the voluntary nature of this module, the score will be considered only if it benefits the final result of the teacher's portfolio. The evaluation of Module 3 will in no case harm the outcome of the teacher score. If the teacher has a low score on Module 3, the portfolio final result will be based only on Modules 1 and 2. Therefore, the teacher final score in the portfolio is calculated two times: first, taking into consideration the score for the three modules with the respective weights showed in Table 1.2 for three modules; second, a final score only with the scores in the indicators from Module 1 and Module 2 (with the respective weights for two modules). Both results are compared and the higher score of these two calculations will be taken as the final result in the teacher portfolio.

Table 3.6 presents the indicators for Module 3, with the performance standards for a competent teacher.

Table 3.6.
*Indicators and Performance Standards for Module 3: Collaborative Work*

| Indicators | Performance Standards for a Competent Teacher |
|---|---|
| 5.1. Collaborative work suitability | The teacher has participated in a collaborative work experience that allowed them to address a problem, need, or important issue in their teaching and that ultimately aimed to improve student learning. |
| 5.2. Quality of professional dialogue | In the collaborative work experience, the teacher participated in a dialogue focused on pedagogical issues, reflecting together on the needs of their students. This reflection was shown through the questions asked to understand the complexity of the problem, the arguments and counterarguments they presented, and the interpretations and explanations that were given. |
| 5.3. Value of collaborative work in professional development | From the collaborative work experience, the teacher learned new techniques that enriched their teaching practice. The teacher recognized how working with the rest of the participants helped them to achieve their new knowledge, by giving them the possibility to question their own pre-conceived notions and teaching practices. |
| 5.4. Reflection on the impact of the collaborative work experience | The teacher analyzed the results of the collaborative work experience and realized how these results impacted the students' learning. |

*Note.* Translated and adapted from www.docentemas.cl

Each one of the portfolio indicators is evaluated using a rubric, and the score is converted to one of the four achievement categories: Unsatisfactory, Basic, Competent, and Outstanding. The weight of each indicator is the same, with the exception of four aspects from Module 2 that have a higher weight: clear explanations, questions and

activities, student feedback, and curricular emphasis on the subject. The reason for the higher weight of these four indicators is due to the fact that they are considered to be essential aspects of effective teaching. Another variable in the weighting of each indicator is whether or not Module 3 is included. If Module 3 is a part of the total score, the percentage of each indicator fluctuates between 4% and 9% (20 indicators). If Module 3 is not included, the percentage of each indicator fluctuates between 5% and 10% (16 indicators).

## Procedures

***Evaluation process***. Each year, the Ministry of Education does a teacher evaluation process for all of the teachers from municipal schools who have not been evaluated or were evaluated for years before. The duration of the evaluation process is 12 weeks, in which teachers have to complete the four assessment instruments, including the development of their portfolio. Thus, teachers have 12 weeks to prepare the portfolio, normally from the beginning of August to the end of October[12].

The entire portfolio, with the exception of the video recorded class, must be uploaded to the web site specially created for this purpose. For the portfolio preparation, teachers receive a manual that helps them to develop each one of the module indicators. As a result of the clear manual instruction for the portfolio process, all teachers are expected to follow the standardized format. This manual also contains a specification of the Good Teacher Framework (GTF) descriptors that will be considered when evaluating each indicator, making the evaluation criteria explicit to the teacher.

[12] The school year in Chile runs from the beginning of March to the middle of December.

The video recorded class of Module 2 is carried out by a trained cameraman provided by the Teaching Professional Performance Evaluation System, at no cost to the teacher. The cameramen are taught how to ensure that the filming is audible and that there is adequate image quality. Thanks to this training, the video recorded gives a clear impression of what happened in the class, reflecting the interaction between the teacher and students in a standardized way.

***Correction process***. The portfolio correction process takes about two months, normally between the months of November and January, in different universities throughout Chile. These universities create correction centers to carry out the correction process, which requires different resources, such as the physical space, computers, supervisors, and correctors. The hiring of the correctors seeks a specific professional profile: teachers who have five years of classroom experience in the level and subject of the portfolio they correct, who have a professional teaching certificate, and who pass the selection test for this process (Docentemas, 2020).

The correctors undergo a training process from MIDE UC measurement specialists, focused on the correction rubric content, structure, and application. The rubrics used for the evaluation are based on the GTF and they are categorized into four possible levels of performance: Unsatisfactory, Basic, Competent, and Outstanding. The correctors evaluate each piece of the portfolio individually by indicator. For each one of the indicators, there is a specific rubric with a complete description of the expected teacher performance for each one of the performance level category. Thus, each indicator is rated in one of the four possible levels with a score from 1 to 4 (Docentemas, 2020).

In order to assure the quality of the rating process, all correctors are trained during 30 hours, to get the knowledge of the scoring rubrics and learn to apply it using the same practice portfolios, thus their performance is being monitored permanently during that period (Taut et al., 2012). This period of training is called the calibration process, and it serves to verify that all the quality parameters established for the correction are adequately met (Docentemas, 2020). During all the correction processes, there are professionals in charge of supervising the evaluation performance. They receive a 40-hour training in order to be prepared for their supervision (Taut et al., 2012).

During the correction process, 25% of the portfolios are randomly selected for double correction for each subject taught and grade level (Taut et al., 2012). The same module is reviewed by two correctors in a "blind" form (ignoring each of the scores assigned by the other evaluator; Docentemas, 2020). If the two raters differ substantially (more than one point of difference), then the supervisor functions as a third rater who resolves the discrepancies, and divergent raters are retrained on the use of the rubric (Taut et al., 2012). This process allows for the detection of differences in the application of the evaluation rubrics (Docentemas, 2020). Another mechanism that is used to ensure the reliability of the process is that every Monday during the evaluation process, all raters complete a group scoring session with their supervisors (Taut et al., 2012).

MIDE UC uses a software program that allows the MIDE UC technical team to monitor the correction process. The specialists verify through the computer system that the process is carried out in compliance with all the established quality protocols. Therefore, the MIDE UC technical team supervises local centers, guarantees the standardization of the correction process in the centers (Docentemas, 2020).

## Analytic Strategy

In an attempt to find answers to the main aim of this dissertation, which is to contribute to the body of research on the evidence of the validity of the Chilean Teacher Evaluation System, I assessed different evidence of validity, focusing specifically on the portfolio. Validity has been understood in the present research as the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests (AERA et al., 2014). Thus, in order to contribute to the evidence of the validity of the Chilean Teacher Evaluation System, I considered the purpose and interpretation of the portfolio results.

For the portfolio used to evaluate the teachers in the Chilean Teacher Evaluation System, two broad purposes have been recognized: providing formative data on individual teachers to improve their practice, and providing summative data to support individual teacher reward and sanctions (Taut et al., 2012). In the previous research carried out by MIDE UC researchers, they took into account these two evaluation purposes, making an important contribution of validity evidence to the Chilean Teacher Evaluation System, making it more valid and relevant for all assessment users. In the present research, I also considered the formative and summative purposes of the Chilean Teacher Evaluation System.

Taking into consideration first the summative purpose of the portfolio used in the Chilean Teacher Evaluation System, one objective was to assess the portfolio internal structure in order to answer the question: does the portfolio scores represent the different aspects of teacher quality as has been declared by the Chilean Teacher Evaluation System?. This question was already studied in previous research (Taut et al., 2012).

However, I extended it taking into consideration that the current portfolio structure (3 modules) has not been validated yet.

Portfolio Module 3 was recently added to the instrument, with the idea of emphasizing collaborative work as part of teacher quality, which is part of domain D from the GTF. Collaborative work measures the interaction, exchange of ideas, and shared reflection between the teacher and their peers. As was previously described, the Module 3 score is only taken into account for the final portfolio score when it benefits the teacher. Therefore, in the portfolio final score it is possible to observe a group of teachers whose collaborative work as an indicator of teacher quality has been included, and another group whose collaborative work has not been included.

The **first aim of this dissertation** was to assess the new structure of the portfolio that was first used in the Chilean Teacher Evaluation System from 2016 on. I also assessed the structure across two different subgroups: teachers whose Module 3 evaluation was taken into account for their final portfolio score, and those teachers whose Module 3 was not taken into account for their final score.

Factor analysis is one of the most common methodologies used to evaluate the structure of an evaluation instrument. The fundamental principle of factor analysis is to determine the number and nature of the latent variables or factors that account for the variation and covariation among a set of observed measures. The observed measures or indicators are intercorrelated because they share a common factor model (Brown, 2015).

The two main types of analyses based on the common factor model are exploratory factor analysis (EFA) and confirmatory factor analysis (CFA). Both analysis aim to reproduce the observed relationship among a group of indicators with a smaller set

of latent variables. However, they differ by the number and nature of a priori specifications and restrictions on the factor model. EFA is a data-driven approach, therefore, no specifications are made in regard to the numbers of factors or the pattern of the relationship between the factors and the indicators. EFA has been used as an exploratory technique to determine the appropriate number of common factors and the measured variables are indicators of the latent dimensions. In contrast, in CFA the researcher specifies the number of factors and the pattern in advance, as well as other parameters. This prespecified solution is evaluated in terms of how well it reproduces the covariance matrix of the measured variables. CFA requires a strong conceptual foundation to guide the specification and evaluation of the factor model (Brown, 2015).

A relatively recent approach implemented in M*plus* program is Exploratory Confirmatory Factor Analysis (ECFA). The main advantage of the ECFA model over the existing approaches explained above is that it seamlessly incorporates the EFA and CFA models. In most applications with multiple factors, the EFA is used to discover and formulate factors. Usually, EFA is followed by an ad-hoc procedure that mimics the EFA factor definitions in an SEM model with a CFA measurement specification. The ECFA approach accomplishes this task in a one-step approach and thus it is a simpler approach. Additionally, ECFA is more accurate because it avoids potential pitfalls due to the challenging EFA to CFA conversion by estimating the measurement and structure model parts simultaneously. Therefore, compared to EFA or CFA, ECFA offers a greater amount of flexibility in the case of model uncertainty (Asparouhov & Muthén, 2009).

In order to assess the evidence of validity based on the internal structure of the portfolio for the present research, I used the ECFA approach to determine the number of

factors underlying the **20 portfolio indicators** for the whole sample of teachers evaluated and for the subgroup of teachers whose Module 3 score was taken into account for their final portfolio score. **16 portfolio indicators** were used for the subgroup of teachers whose Module 3 score was not taken into account for their final score. ECFA was used with oblique Geomin rotation, evaluating five separate models that represented factor solutions from one to five dimensions, for each one of the samples analyzed (whole sample, teachers with M3 score, and teachers without M3 score).

Within the results, I used different fit indexes in order to evaluate the data fitting in each one of the models tested. Three commonly used fit indexes are:

1. **The Comparative Fit Index (CFI)**. An incremental Fit Index, which assesses the degree to which the tested model is superior to an alternative model in reproducing the observed covariance matrix. Therefore, the larger the number, the better the model fit, since larger values indicate greater improvement of model fit over an alternative model (Chen, 2007). CFI values close to 0.95 or greater indicate a reasonably good fit between the target model and the observed data (Hu & Bentler, 1995).

2. **Root mean square error of approximation (RMSEA)**. An absolute Fit Index, that assesses the degree to which the model-implied covariance matrix matches the observed covariance matrix. The smaller the number, the better the model fit. A value of 0 indicates an optimal fit, and increasing values indicate departure of the covariance matrix from the observed matrix (Chen, 2007). RMSEA values close to 0.06 or below indicate a reasonably good fit between the target model and the observed data (Hu & Bentler, 1995).

3. **Standardized root mean square residual (SRMR)**. Also, an absolute Fit Index. It is a measure of the average of the standardized residuals between the observed and model-implied covariance matrixes (Bentler, 1995). The smaller the number, the better the model fit. (Chen, 2007). SRMR values that are close to 0.08 or below indicate a reasonably good fit between the target model and the observed data (Hu & Bentler, 1995).

Taking into consideration the comparative fit indices described above, I evaluated the fit of the five separate models that represented factor solutions from one to five dimensions, for each one of the samples analyzed in the present research. I considered the cutoff points described previously to evaluate the goodness of fit. Later, Modification Index (MI) results were used to identify which parameters, if freely estimated, could contribute to a significant drop in the chi-square statistics and could improve the fitting indexes for all of the five proposed factor solution models. Finally, I assessed the structure of the loading using the rule of thumb that ignores loadings less than 0.3.

Also, considering the summative purpose of the portfolio, a **second objective was to determine if the same evaluation instrument (portfolio) correctly distinguishes overall teacher quality for teachers who teach in different contexts or settings.** As already indicated, the portfolio used in teacher evaluation is a unique instrument that evaluates teachers from different teaching levels (with the only exception of technical professional teachers) and location (rural/urban). However, when these variables are considered, it is possible to observe differences in the teaching circumstances.

Research has been showing that the context in which teachers work can affect the evaluation of their teaching performance. Therefore, the quality of teaching measured by

the results of the educational process is influenced by the instructional context (Bryk et al., 2012; Darling-Hammond, 2010). Teachers should be evaluated considering the institution, the student population, and the resources with which they work (Everson et al., 2013).

Factors such as school location could be considered as an adverse context that impedes teachers' performance. In the context of Chilean rural teachers, working in multi-grade classrooms, areas with difficult access, vulnerability, and high poverty cases are characteristics that they have to face during the teacher evaluation process. This puts them in a different position from their urban pairs, and this difference is not taken into account in the evaluation system (Castillo-Miranda et al., 2017). Thus, the results that indicate teachers from rural schools having a lower score than the urban ones, could be due to the conditions that exist based on the context (Colegio de Profesores de Chile A.G., 2016).

Evidence related to different practices from teachers across different grade levels has also been presented by the research (Vartuli, 1999). Some research has shown that elementary teachers obtain better rating evaluation compared to middle school teachers (Harris & Sass, 2011). Thus, the level in which teachers teach could impact their results on the portfolio evaluation.

Therefore, **the second aim of the present dissertation** was to provide validity evidence that determined if the portfolio factor structures were invariant across subgroups, such as different teaching levels and school location (rural/urban). In seeking evidence of multigroup invariance, researchers are interested in finding the answer to questions such as whether or not the items comprising a particular measuring instrument

operate equivalently across different populations. In other words, is the measurement model group-invariant? (Byrne, 2012). An instrument has measurement invariance across groups if subjects with identical levels of the latent construct have the same expected raw-score on the measure (Drasgow & Kanfer, 1985). When measurement invariance is established, observed mean differences can be attributed to differences in the underlying construct between groups. On the other hand, if measurement invariance cannot be assumed, observed mean differences may be due to the different relations between the latent constructs and scores (Hirschfeld & Brachel, 2014).

For the Chilean Teacher Evaluation System, the same portfolio has been used for all teachers to determine if they have to be reevaluated the following year (an unsatisfactory result), or if they have to participate in specific professional development plans to address their weakness (a basic result), or if they can progress through a better level in their professional career (an outstanding or a competent result). However, can the observed differences in the teachers' results be attributed to differences in the teacher quality construct evaluated? Measurement invariance was used to answer the question of whether the portfolio measures the same construct across different teacher groups.

Measurement invariance is defined as "the mathematical equality of corresponding measurement parameters for a given factorially defined construct (i.e., the loadings and intercepts of a construct's multiple manifest indicators) across two or more groups" (Little, 1997, p. 55). Portfolio measurement invariance for different teacher groups was tested within the framework of multigroup Confirmatory Factor Analysis (CFA) modeling using procedures outlined by Byrne (2012). Analyses were conducted using the M*plus* 8 program.

Testing for factorial invariance encompasses a series of steps that build upon one another, with a series of model comparisons that define more and more stringent equality constraints (Byrne, 2012; Cheung & Rensvold, 2002). There are four levels of factorial invariance: configural invariance, weak factorial invariance, strong factorial invariance, and strict factorial invariance (not recommended because the criterion is too strict and hard to put into practice).

***Baseline Model***. At the beginning of the analysis, a baseline model in which the loading patterns are similar in all groups but the loading patterns may vary, has to be fit (Hirschfeld & Brachel, 2014). For the present dissertation, I first established a well-fitting baseline model for each group (rural/urban, six different teaching levels). Therefore, once a baseline model was identified across teaching levels taught, and across school location, I tested the equivalence of this model, imposing a series of increasingly stringent between-group constraints in several nested models that are described below. A robust weighted least square estimator (WLSMV) was used, as it is the appropriate estimation for categorical ordered data.

***Configural Invariance***. The first model specified configural invariance, meaning that the same factor structure is invariant across groups. The baseline model has a good fit and the same loadings are significant in all groups (Hirschfeld & Brachel, 2014). For the present research, the same items loaded onto the same factors across groups, were estimated simultaneously within the two groups for school location and the six teaching level groups. I use chi-square statistics that indicate when a relevant deviation of the data from the model is significant. However, chi-square can be affected by the large sample size (Dimitrov, 2010). Thus, other fit measures such as the root mean square error of

approximation (RMSEA), the standardized root mean square residual (SRMR), and the comparative fit index (CFI) to evaluate the goodness of fit have been used. In order to evaluate the fit index supporting for a reasonably good fit, the cutoff criteria used for each index was: 1) CFI value close to 0.95 or greater; 2) RMSEA value close to 0.06 or below; and 3) SRMR value close to 0.08 or below (Hu & Bentler, 1995).

When the goodness of fit parameters supported a reasonably good fit, this indicated configural invariance. If it does have configural invariance, the configural model becomes the model with which subsequent models are compared (Bowen & Masa, 2015).

***Weak factorial invariance***. When a configural invariance model is supported, the second step is to test for weak invariance, in which the factor loadings are constrained to be equal to the data and the fit of this model is compared to the configural model. Weak invariance is supported if the fit of the metric invariance model is not substantially worse than the previous model (Hirschfeld & Brachel, 2014). For this research, I proceeded to test Model 2: weak factorial invariance, in which corresponding factor loadings were equivalent across groups.

In order to test if the model is not substantially worse than the previous one, there are different decision rules. Initial studies used a chi-square test to decide if the next model increases in fit substantially (Byrne, 2012). However, as has been mentioned previously, chi-square can be affected by the large sample size (Dimitrov, 2010). Thus, a number of goodness of fit indexes are used to judge the absolute model fit to support invariance, in addition to a chi-square test (Chen, 2007; Cheung & Rensvold, 2002). Through a Monte Carlo simulation study, Cheung & Rensvold (2002) examined changes

in goodness of fit indexes caused by invariance constraints across groups. The results indicate that only CFI differences were not affected by the specification accuracy in the overall model. Therefore, they recommended that researchers report ΔCFI for testing with invariance, which should not be retained when there is a decrease of .01 or larger. That is, a negative ΔCFI value equal to or lower than −.01 (e.g., ΔCFI = −.02) would indicate a lack of invariance (Cheung & Rensvold, 2002). Likewise, Chen (2007) also based on Monte Carlo Studies, proposes cutoff points based on the three routinely used fit indexes (i.e., CFI, RMSEA, and SRMR), recommended evaluating invariance at the three commonly tested levels. The author indicated that the cutoff point when the sample size is adequate (total $N > 300$ ), for testing loading invariance (weak factorial invariance), should be a change equal to or lower than -.010 in CFI, supplemented by a change equal to or higher than .015 in RMSEA, or a change equal to or higher than .030 in SRMR would indicate noninvariance (Chen, 2007).

In order to check for invariance in the weak model, first I evaluated the fit of Model 2 using the chi-square DIFFTEST, comparing it with the configural model. RMSEA, CFI, and SRMR changes were also used to evaluate the progressive factorial invariance between the configural and the weak factorial invariance models, following the approach used by Chen (2007), with the cutoff points indicated above.

***Strong factorial invariances***. In addition to invariant item factor loadings and the same patterns of item loading on each factor, strong factorial invariance is used to test factorial invariance for the item intercepts constrained to be equal. In the case of ordinal indicators, strong factorial invariance is evaluated by the presence of invariant thresholds

for each indicator that describes at which level of the latent variable a specific category is chosen, instead of intercepts (Hirschfeld & Brachel, 2014).

Strong factorial invariance implies that differences in scale scores are due to the differences in true levels of the underlying construct. Strong variance is supported if the fit of the scalar invariance model is not substantially worse than the fit of the weak invariance model (Hirschfeld & Brachel, 2014). For this research, I proceed to test Model 3: strong factorial invariance, in which invariant thresholds for each indicator were equivalent across groups.

The chi-square DIFFTEST with the weak model RMSEA, CFI, and SRMR changes were also used to evaluate the progressive factorial invariance between the weak and strong factorial invariance models. Following Chen (2007), the cutoff points recommended for evaluating invariance at intercept (threshold) levels are a change of equal to or lower than -.010 in CFI, supplemented by a change of equal to or higher than .015 in RMSEA, or a change equal to or higher than .010 in SRMR, that would indicate noninvariance (Chen, 2007). Figure 3.1 shows the decision-making process in order to indicate measurement invariance.

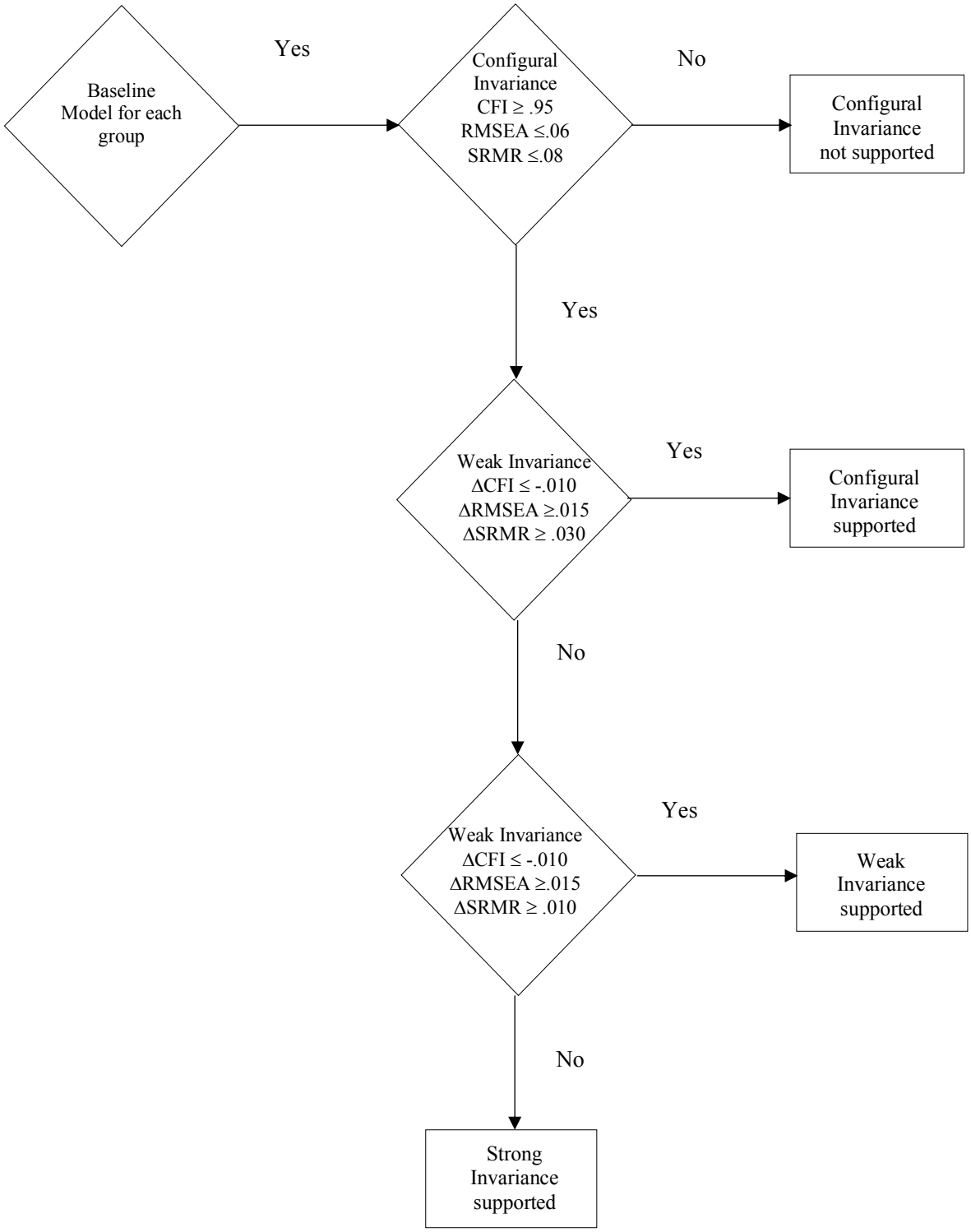


*Figure 3.1.* Selection decisions for Multiple Invariance support

**A third aim of the present dissertation** is related to providing evidence for the portfolio final score calculated by theoretical weight assigned to each one of the portfolio indicators with the empirical data. This objective was also aligned with evidence of the portfolio related to its summative purpose.

As has been explained previously, the portfolio final score is calculated with the score obtained by the teacher in the evaluation of each one of the indicators, but considering the specific weight of each one. This weighted final score has been changed since the portfolio modifications in 2016, because before the change each indicator in the portfolio weighed the same. The reason behind this modification was because they have been considered to be essential aspects of effective teaching. Therefore, the weight of each indicator is the same, with the exception of four aspects from Module 2 that have a higher weight: clear explanations, questions and activities, student feedback, and curricular emphasis on the subject. Another variable in the weighting of each indicator is whether or not Module 3 is included. If Module 3 is a part of the total score, the percentage of each indicator fluctuates between 4% and 9% (20 indicators). If Module 3 is not included, the percentage of each indicator fluctuates between 5% and 10% (16 indicators).

In order to assess the empirical weights, I used a weighted sum score method. With this method, the sum score can be obtained when the factor loading of each item is multiplied to the scaled score for each item before summing. One advantage of the weighted sum score method is that items with the highest loading on the factor have the largest effect on the factor score (Distefano et al., 2009). Later, in order to compare the

scores calculated by the theoretical weighted score for the portfolio and the weighted sum score, I compared both scores using a paired *t*-test for significant differences.

Finally, the **fourth aim of this dissertation** was related to the formative purpose of the portfolio. One of the main objectives of the portfolio in the Chilean Teacher Evaluation System is the promotion of teacher improvement through the evaluation of their teaching practices and professional development. From the portfolio evaluation, each teacher receives a report with a complete feedback report on different aspects of their practice. From this information, they can reflect individually and with their colleagues in order to make efforts to improve on relevant aspects of their practice (Docentemas, 2020). In order to contribute to the portfolio formative purpose, this objective evaluates validity evidence that supports the interpretation and use of portfolio subscores.

The information provided by the portfolio resulted in an aggregated final score. However, more detailed information would provide more evidence to the teacher about their strengths and weaknesses. For aim four, the portfolio subscores were proposed to evaluate whether or not portfolio subscores have added value over the total score. The portfolio is conformed of different domains and modules. Thus, the portfolio subscores that were evaluated by these objectives were done first at the module level, and second at the domain level, specifically within the three domains that compose Module 1.

Given the importance of subscore reporting, the quality of subscores must be assessed to avoid inaccurate information at the subscore level. Inaccurate subscore reporting leads to incorrect instructional and remedial decisions, resulting in needless time and effort spent (Sinharay & Haberman, 2008). One possible reason for inaccurate

subscore reporting is that the subscores reported are less different from one another, so they can become redundant and possibly misleading (Feinberg & Jurich, 2017). Thus, the subscores must report additional information from the test that would otherwise not be reported (Feinberg & Wainer, 2014).

Researchers have developed different approaches for evaluating whether **subscores have added value** over the total score. One of the approaches used by Haberman (2008) is the **proportional reduction of the mean squared error (PRMSE).** This method is based on classical test theory (CTT) and can be used for evaluating the precision of subscores and total scores as predictors. The logic behind PRMSE is based on two conditions: observed subscores are most likely to have value if they have relatively high reliability by themselves and if the true subscore and true total score have only a moderate correlation (Haberman, 2008). The more orthogonal the subscore is to the rest of the test, the greater the value (Feinberg & Wainer, 2014). PRMSE provides a marginal value measurement that the subscore adds to the test. Thus, this method has been used to determine whether and how to report subscores.

This approach assumes that a reported subscore is intended to be an estimate of the true subscore (*St*) (Sinharay, 2010). The estimate of the true subscores is:

$S_s = \bar{S} + \alpha(S - \bar{S})$, where $\bar{S}$ is the average subscore for the sample of examinees and $\alpha$ is the reliability of the subscore;

$S_x = \bar{S} + C(x - \bar{x})$, based on the observed total score, where $\bar{x}$ is the average total score and *C* is a constant that depends on the reliabilities and standard deviations of the subscore, and the total score and the correlation between the subscores;

$S_{sx} = \bar{S} + a(S - \bar{S}) + b(x - \bar{x})$, the weighted average of the observed subscore and the observed total score, where *a* and *b* are constants that depend on the reliability and standard deviations of the subscore, and the total score and the correlation between the subscores.

In order to compare the performances of *Ss, Sx,* and *Ssx* as estimates of *St*, Haberman (2008), suggests using the proportional reduction in mean squared error (PRMSE). A subscore has added value over the total score only if PRMSE*s* (*Ss*) is larger than PRMSE*x* (*Sx*), because in that case, the subscore will provide more accurate diagnostic information than the observed total score (Sinharay, 2010). The formula developed by Haberman (2008) is an important contribution because it helps determine which subscores on a test deserve being reported and which ones should not be reported (Feinberg & Wainer, 2014).

Under the same logic of PRMSE, Feinberg and Wainer (2014) refined the Haberman (2008) method by presenting the *PRMSEs* as a ratio that they called value-added ratio (VAR). The VAR is presented in a simple equation to approximate *PRMSEs* values:

$$\left(\frac{PRMSE_s}{PRMSE_x}\right) = VAR \approx 1.15 + 0.5 \times r_1 - 0.67 \times r_2$$

where $r_1$ is subscore reliability and $r_2$ is the disattenuated correlation between the subscore and the score remainder (composed of the remaining items on the test not included in the subset from which the subscore was computed; Feinberg & Jurich, 2017). The disattenuated correlation refers to the application of a correction for attenuation to an observed correlation that takes into account the unreliability of the variables (DeVellis,

2017). It is calculated dividing the raw correlation between the subscore (*x*) and remainder of the test (*y*) by the square root of the product of their reliabilities (Feinberg & Jurich, 2017):

$$r_2 = \frac{r_{xy}}{\sqrt{r_{xx} r_{yy}}}$$

There is a broad agreement in the literature that the value for VAR needs to be >1 for a subscore to be worth reporting (Feinberg & Wainer, 2014; Haberman, 2008). In order to determine if the subscore evaluated yields an added value over the total score, the results for the VAR equations should be greater than one.

The extent to which VAR approximations is an accurate measure of the actual ratio of *PRMSEs* is up for debate (Sinharay et al., 2015). However, the results presented by the authors indicate that both approaches, PRMSE and VAR, yield the same supporting evidence of subscore, leading to the same decisions on whether or not to report subscores (Feinberg & Wainer, 2015). For the present research, I used the Feinberg and Wainer (2014) approach.

# Chapter 4: Results

This chapter reports the results of this study, contributing to the body of research on the evidence of the validity of the portfolio used by the Chilean Teacher Evaluation System. The results were based on the data of the teacher evaluation carried out in 2017 that evaluated 24,251 teachers who worked in the Chilean municipal schools. The teachers evaluated in 2017 taught different grade levels, subjects, and different school locations (urban/rural). They were evaluated using the same portfolio, regardless of those differences, with the exception of the technical/professional education teachers, for whom the portfolio contained three different indicators in order to align the teacher evaluation with the technical/professional teachers' work. Considering these differences, technical education teachers were not part of the present study. Therefore, the results of the present section are based on the data of 21,982 teachers evaluated in 2017.

For each aim of the study, I present the results from statistical analyses including descriptive statistics.

## Aim 1: Assessing the structure of the portfolio across two different subgroups

The first aim of this study assessed the structure of the portfolio taking into consideration two different subgroups of teachers: those whose Module 3 was included in their final score and those teachers whose Module 3 was not taken into account for the final portfolio score. In order to answer aim 1, I used Exploratory Confirmatory Factor Analysis (ECFA) to assess the structure of the portfolio for: the whole sample of teachers evaluated, the teachers with Module 3 in their final score, and the teachers without Module 3 in their final score. As described previously, Module 3 has been included in the portfolio since 2016, but the score is only taken into account as part of the final portfolio score when it benefits the teacher. For the 2017 evaluation results, 52.33% of the teachers' final score (*N*=11,216) included the Module 3 evaluation, and 47.67% of the teachers' final score (*N*=10,216; missing = 550) did not. From the group of teachers in which the Module 3 portfolio was not included in their final score, 34% did not even submit any evidence to be evaluated (*N*=3,479).

### Descriptive Statistics for teachers with M3 and without M3

Table 4.1 describes the frequencies and percentages of the teachers' performance level rating from unsatisfactory (1) to outstanding (4), for each indicator of the seven portfolio indicators that are part of Module 1. The table indicates the results for the whole sample of teachers evaluated, teachers with the Module 3 evaluation, and those teachers whose Module 3 was not taken into account. As we can see in Table 4.1, for the whole sample of teachers, indicators that are related to planning, that evaluate a series of written documents of pedagogical materials submitted by the teachers for their evaluation

(indicators 1.1 and 1.2), presented the highest percentage of teachers rated in the highest category of teacher performance (competent or outstanding). Conversely, the indicators that are related to assessment (indicators 2.1, 2.2, and 2.3) presented the lowest percentage of teachers evaluated as competent or higher. The indicators associated with reflection (3.1 and 3.2) had mixed results, with one indicator showing a higher percentage of teachers evaluated as basic or lower (indicator 3.1), and the other indicator showing a higher percentage of teachers rated as competent or outstanding.

Therefore, the results for Module 1 indicate that all of the teachers evaluated in the year 2017 showed better results on the teacher quality variables related to the planning dimension. The planning dimension evaluates aspects of teacher quality as the formation of learning objectives and the relationship between activities and objectives. On the other hand, teachers showed lower results on the variables of teacher quality related to the assessment dimension, the dimension that evaluated teacher use of evaluation and rubrics for correction, the relationship between assessment and objectives, and the analysis and use of assessment results.

When I observe just the teachers whose Module 3 was considered in their final score, the results of the percentage of teachers evaluated as basic or competent were similar and consistent with the results of the whole sample of teachers evaluated, presented above. However, for the teachers that had Module 3 as part of their final score, the percentage of teachers that were rated as unsatisfactory was lower compared to the whole sample of teachers across the seven indicators. On the other hand, the percentage of teachers who were evaluated as outstanding was higher for teachers with Module 3 included in their evaluation. It is possible to observe an opposite trend between the whole

sample and the teachers whose Module 3 was not part of their final score. There was a higher proportion of unsatisfactory teachers and a lower proportion of outstanding teachers when compared with the whole sample.

With respect to the differences between the teachers whose Module 3 was taken into account for the final score, and those teachers whose Module 3 was not included, the proportion of teachers evaluated in each category score was significantly different for both groups in all of the seven indicators from Module 1. In all of the indicators, the group of teachers that were evaluated including Module 3 yielded a higher proportion of teachers as competent or higher[13] than the teachers without Module 3 (Ind.1.1: 91.04% vs. 89.24%, $\chi^2_{(3)}=71.61$, $p < 0.01$; Ind.1.2: 64.94% vs. 62.05%, $\chi^2_{(3)}=32.8$, $p < 0.01$; Ind.2.1: 39.19% vs. 36.94%, $\chi^2_{(3)}=48.58$, $p < 0.01$; Ind.2.2: 49.92% vs. 47.42%, $\chi^2_{(3)}=82.72$, $p < 0.01$; Ind.2.3: 50.6% vs. 40.95%, $\chi^2_{(3)}=288.1$, $p < 0.01$; Ind.3.1: 50.52% vs. 39.77%, $\chi^2_{(3)}=363.0$, $p < 0.01$; Ind.3.2: 71.09% vs. 58.31%, $\chi^2_{(3)}=432.15$, $p < 0.01$).

[13] From here on, for all the chi-square comparison results presented the comparison percentages between groups considering the sum of the percentages of teachers rated as competent and outstanding.

Table 4.1.

*Frequencies and percentages for the 7 Module 1 portfolio indicators for all teachers, and teachers whose M3 was and was not taken into consideration for the final score*

| Indicator | All teachers | | Teachers with M3 results | | Teachers without M3 results | |
|---|---|---|---|---|---|---|
| | Frequency | Percent | Frequency | Percent | Frequency | Percent |
| 1.1.Formation of learning objectives | | | *** | | *** | |
| Unsatisfactory | 468 | 2.18 | 185 | 1.65 | 283 | 2.77 |
| Basic | 1,636 | 7.63 | 820 | 7.31 | 816 | 7.99 |
| Competent | 16,390 | 76.49 | 8,510 | 75.89 | 7,880 | 77.15 |
| Outstanding | 2,934 | 13.69 | 1,699 | 15.15 | 1,235 | 12.09 |
| 1.2.Relationship between activities and objectives | | | *** | | *** | |
| Unsatisfactory | 3,027 | 14.13 | 1,473 | 13.14 | 1,554 | 15.21 |
| Basic | 4,780 | 22.31 | 2,458 | 21.92 | 2,322 | 22.73 |
| Competent | 13,144 | 62.34 | 6,998 | 62.40 | 6,146 | 60.17 |
| Outstanding | 477 | 2.23 | 285 | 2.54 | 192 | 1.88 |
| 2.1. Evaluation and rubrics used for correction | | | *** | | *** | |
| Unsatisfactory | 3,510 | 16.55 | 1,681 | 15.06 | 1,829 | 18.21 |
| Basic | 9,611 | 45.32 | 5,106 | 45.75 | 4,505 | 44.85 |
| Competent | 7,840 | 36.97 | 4,219 | 37.80 | 3,621 | 36.05 |
| Outstanding | 244 | 1.15 | 155 | 1.39 | 89 | 0.89 |
| 2.2.Relationship between assessment and objectives | | | *** | | *** | |
| Unsatisfactory | 3,591 | 16.93 | 1,670 | 14.96 | 1,921 | 19.13 |
| Basic | 7,281 | 34.33 | 3,921 | 35.12 | 3,360 | 33.46 |
| Competent | 9,030 | 42.58 | 4,799 | 42.99 | 4,231 | 42.13 |
| Outstanding | 1,305 | 6.15 | 774 | 6.93 | 531 | 5.29 |
| 2.3.Analysis and use of assessment results | | | *** | | *** | |
| Unsatisfactory | 3,942 | 18.64 | 1,674 | 15.01 | 2,268 | 22.69 |
| Basic | 7,468 | 35.31 | 3,835 | 34.39 | 3,633 | 36.35 |
| Competent | 9,168 | 43.35 | 5,280 | 47.34 | 3,888 | 38.90 |
| Outstanding | 569 | 2.69 | 364 | 3.26 | 205 | 2.05 |

*p<0.05; **p<0.01;***p<0.001 (Chi-square test)

Note: Chi-square test with significance levels used for differences between the teachers whose Module 3 was taken into account for the final score, and those teachers whose Module 3 was not included.

| Indicator | All teachers | | Teachers with M3 results | | Teachers without M3 results | |
|---|---|---|---|---|---|---|
| | Frequency | Percent | Frequency | Percent | Frequency | Percent |
| 3.1.Analyses based on students' characteristics | | | *** | | *** | |
| Unsatisfactory | 2,199 | 10.30 | 834 | 7.44 | 1.365 | 13.46 |
| Basic | 9,454 | 44.29 | 4,711 | 42.04 | 4,743 | 46.77 |
| Competent | 9,217 | 43.18 | 5,344 | 47.68 | 3,873 | 38.19 |
| Outstanding | 478 | 2.24 | 318 | 2.84 | 160 | 1.58 |
| 3.2.Use of error for learning | | | *** | | *** | |
| Unsatisfactory | 2,067 | 9.71 | 780 | 6.97 | 1,287 | 12.74 |
| Basic | 5,381 | 25.27 | 2,457 | 21.95 | 2,924 | 28.95 |
| Competent | 13,280 | 62.36 | 7,589 | 67.80 | 5,691 | 56.35 |
| Outstanding | 566 | 2.66 | 368 | 3.29 | 198 | 1.96 |

*p<0.05; **p<0.01;***p<0.001 (Chi-square test)

Note: Chi-square test with significance levels used for differences between the teachers whose Module 3 was taken into account for the final score, and those teachers whose Module 3 was not included.

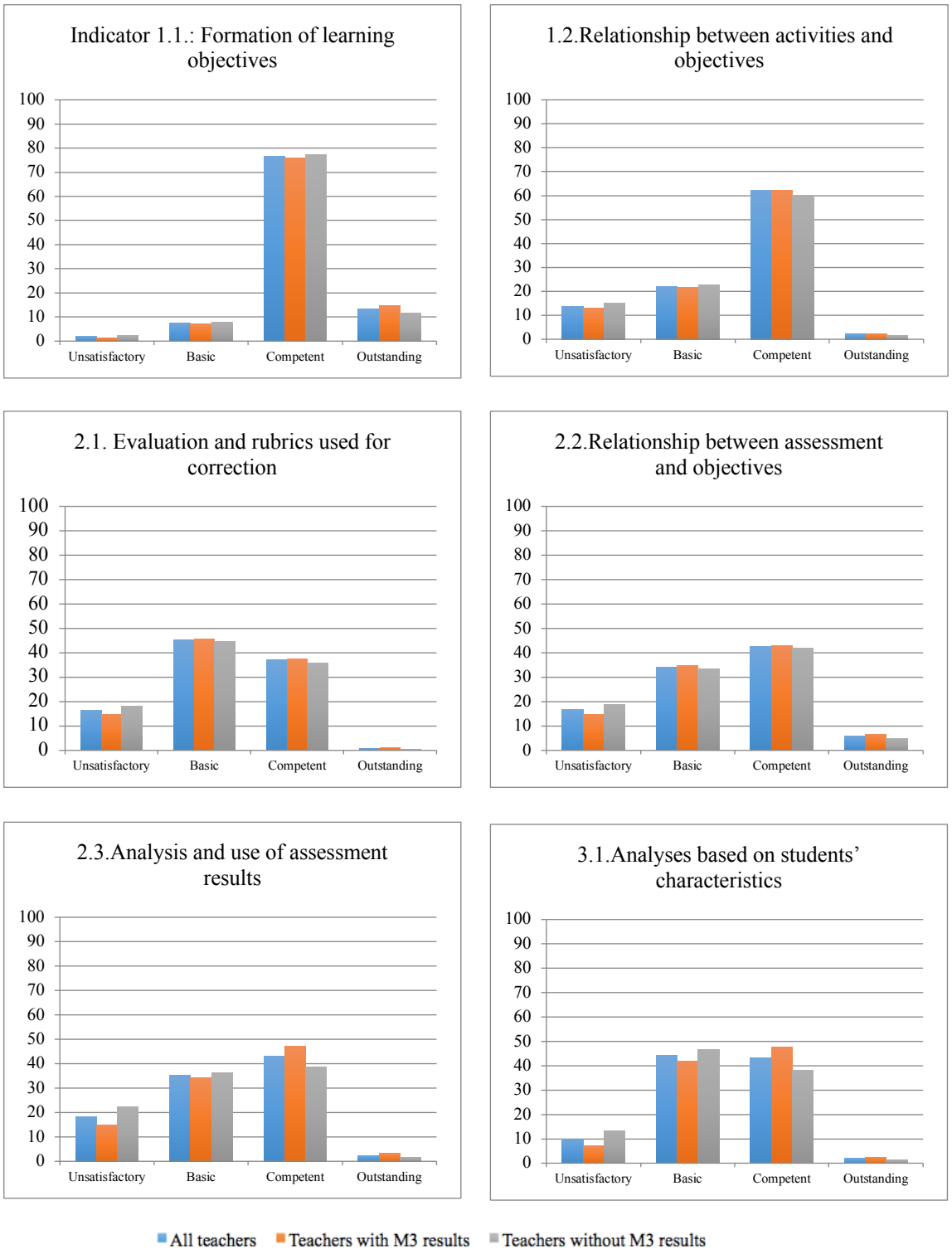


*Figure 4.1.* Percent of responses for the 7 Module 1 portfolio indicators for all teachers, and teachers whose M3 was and was not taken into consideration for the final score.

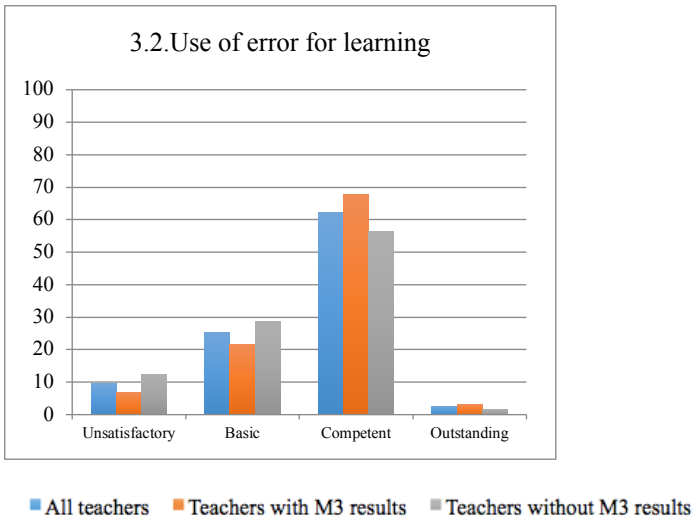


*Figure 4.1.* Percent of responses for the 7 Module 1 portfolio indicators for all teachers, and teachers whose M3 was and was not taken into consideration for the final score.

The frequencies and percentages of the teacher's performance level rating for each one of the nine Module 2 portfolio indicators are presented in Table 4.2. For all of the teachers evaluated, indicators such as *class environment*, *quality of the start of the class, the contribution of the activities to achievement*, and *student encouragement* were rated mostly as competent or outstanding. On the contrary, indicators such as *quality of the end of the class*, *curricular emphasis*, *clear explanations*, *questions and activities*, and *student feedback* were evaluated mainly as basic or unsatisfactory.

Therefore, the results for all teachers evaluated on Module 2 indicate that they showed better results on teacher quality variables related to the classroom environment and the capacity for the teachers to motivate the students to participate in the class. On the other hand, lower results were related to teacher quality variables associated with teacher strategies used in the classroom that promote student learning.

This pattern was similar for both groups of teachers: the ones where module 3 was included and the other one where it was not included. However, there was a small difference in the extreme scores (unsatisfactory and outstanding), which is similar to the outcomes of Module 1.

When I compared the differences between these two groups of teachers, significant differences were found in the proportion of teachers corresponding to each category rate for six of the nine indicators. For those indicators where significant differences were found, the group of teachers with the Module 3 score showed a higher proportion of teachers as competent or outstanding than the teachers without Module 3 (Ind.4.1: 93.83% vs. 93.96%, $\chi^2_{(3)}=17.92$, $p < 0.01$; Ind.4.4: 74.62% vs. 72.19%, $\chi^2_{(3)}=27.46$, $p < 0.01$; Ind.4.7: 38.22% vs. 38.12%, $\chi^2_{(3)}=12.29$, $p < 0.01$; Ind.4.8: 68.92% vs. 68.12%, $\chi^2_{(3)}=16.62$, $p < 0.01$), with the exceptions of indicators 4.6 and 4.9 (Ind.4.6: 34.44% vs. 36.38%, $\chi^2_{(3)}=10.68$, $p = 0.01$; Ind.4.9. 35.34% vs. 33.35%, $\chi^2_{(3)}=12.0$, $p < 0.01$).

Table 4.2.

*Frequencies and percentages for the 9 Module 2 portfolio indicators for all teachers, and teachers whose M3 was and was not taken into consideration for the final score*

| Indicator | All teachers | | Teachers with M3 results | | Teachers without M3 results | |
|---|---|---|---|---|---|---|
| | Frequency | Percent | Frequency | Percent | Frequency | Percent |
| 4.1.Class environment | | | *** | | | |
| Unsatisfactory | 212 | 0.99 | 93 | 0.83 | 119 | 1.17 |
| Basic | 1,078 | 5.03 | 581 | 5.18 | 497 | 4.87 |
| Competent | 19,147 | 89.38 | 10,074 | 89.83 | 9,073 | 88.87 |
| Outstanding | 986 | 4.60 | 466 | 4.16 | 520 | 5.09 |
| 4.2.Quality of the start of class | | | | | | |
| Unsatisfactory | 963 | 4.50 | 482 | 4.30 | 481 | 4.71 |
| Basic | 5,675 | 26.49 | 2,932 | 26.15 | 2,743 | 26.87 |
| Competent | 14,499 | 67.68 | 7,647 | 68.19 | 6,852 | 67.12 |
| Outstanding | 286 | 1.34 | 153 | 1.36 | 133 | 1.30 |
| 4.3.Quality of the end of class | | | | | | |
| Unsatisfactory | 1,680 | 7.84 | 834 | 7.44 | 846 | 8.29 |
| Basic | 9,573 | 44.69 | 5,039 | 44.93 | 4,534 | 44.41 |
| Competent | 9,490 | 44.30 | 4,991 | 44.51 | 4,499 | 44.07 |
| Outstanding | 680 | 3.17 | 350 | 3.12 | 330 | 3.23 |
| 4.4.Contribution of the activities to the achievement of the class objectives | | | *** | | | |
| Unsatisfactory | 2,000 | 9.34 | 948 | 8.46 | 1,052 | 10.31 |
| Basic | 3,683 | 17.20 | 1,897 | 16.92 | 1,786 | 17.50 |
| Competent | 15,381 | 71.81 | 8,193 | 73.07 | 7,188 | 70.42 |
| Outstanding | 355 | 1.66 | 174 | 1.55 | 181 | 1.77 |
| 4.5.Curricular emphasis on the subject | | | | | | |
| Unsatisfactory | 6,249 | 29.17 | 3,342 | 29.80 | 2,907 | 28.47 |
| Basic | 10,129 | 47.28 | 5,304 | 47.30 | 4,825 | 47.26 |
| Competent | 4,855 | 22.66 | 2,469 | 22.02 | 2,386 | 23.37 |
| Outstanding | 190 | 0.89 | 99 | 0.88 | 91 | 0.89 |

*p<0.05; **p<0.01;***p<0.001 (Chi-square test)

Note: Chi-square test with significance levels used for differences between the teachers whose Module 3 was taken into account for the final score, and those teachers whose Module 3 was not included.

| Indicator | All teachers | | Teachers with M3 results | | Teachers without M3 results | |
|---|---|---|---|---|---|---|
| | Frequency | Percent | Frequency | Percent | Frequency | Percent |
| 4.6.Clear explanations | | | | | * | |
| Unsatisfactory | 1,850 | 8.64 | 973 | 8.68 | 877 | 8.59 |
| Basic | 11,997 | 56.00 | 6,379 | 56.88 | 5,618 | 55.03 |
| Competent | 7,415 | 34.61 | 3,788 | 33.78 | 3,627 | 35.53 |
| Outstanding | 161 | 0.75 | 74 | 0.66 | 87 | 0.85 |
| 4.7.Questions and activities | | | ** | | | |
| Unsatisfactory | 201 | 0.98 | 88 | 0.78 | 122 | 1.20 |
| Basic | 13,036 | 60.85 | 6,840 | 61.00 | 6,196 | 60.69 |
| Competent | 7,473 | 34.88 | 3,939 | 35.13 | 3,534 | 34.62 |
| Outstanding | 704 | 3.29 | 347 | 3.09 | 357 | 3.50 |
| 4.8.Encouragement | | | ** | | | |
| Unsatisfactory | 159 | 0.74 | 59 | 0.53 | 100 | 0.98 |
| Basic | 6,582 | 30.72 | 3,427 | 30.56 | 3,155 | 30.90 |
| Competent | 13,540 | 63.20 | 7,144 | 63.71 | 6,396 | 62.65 |
| Outstanding | 1,142 | 5.33 | 584 | 5.21 | 558 | 5.47 |
| 4.9.Student feedback | | | | | ** | |
| Unsatisfactory | 1,973 | 9.21 | 1,041 | 9.28 | 932 | 9.13 |
| Basic | 12,083 | 56.40 | 6,210 | 55.38 | 5,873 | 57.53 |
| Competent | 6,626 | 30.93 | 3,579 | 31.92 | 3,047 | 29.85 |
| Outstanding | 741 | 3.46 | 384 | 3.42 | 357 | 3.50 |

*p<0.05; **p<0.01;***p<0.001 (Chi-square test)

Note: Chi-square test with significance levels used for differences between the teachers whose Module 3 was taken into account for the final score, and those teachers whose Module 3 was not included.

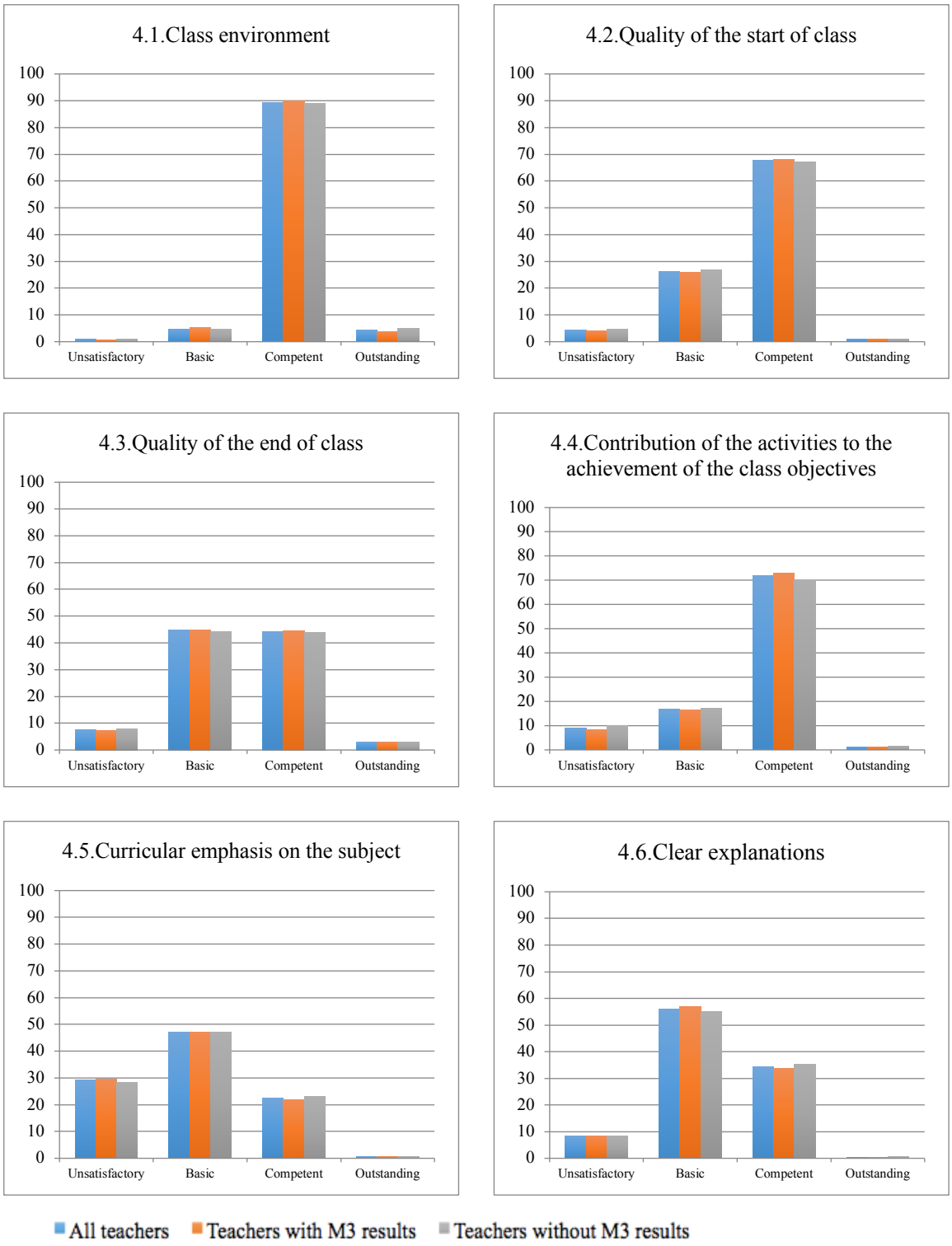


*Figure 4.2.* Percent of responses for the 9 Module 2 portfolio indicators for all teachers, and teachers whose M3 was and was not taken into consideration for the final score.

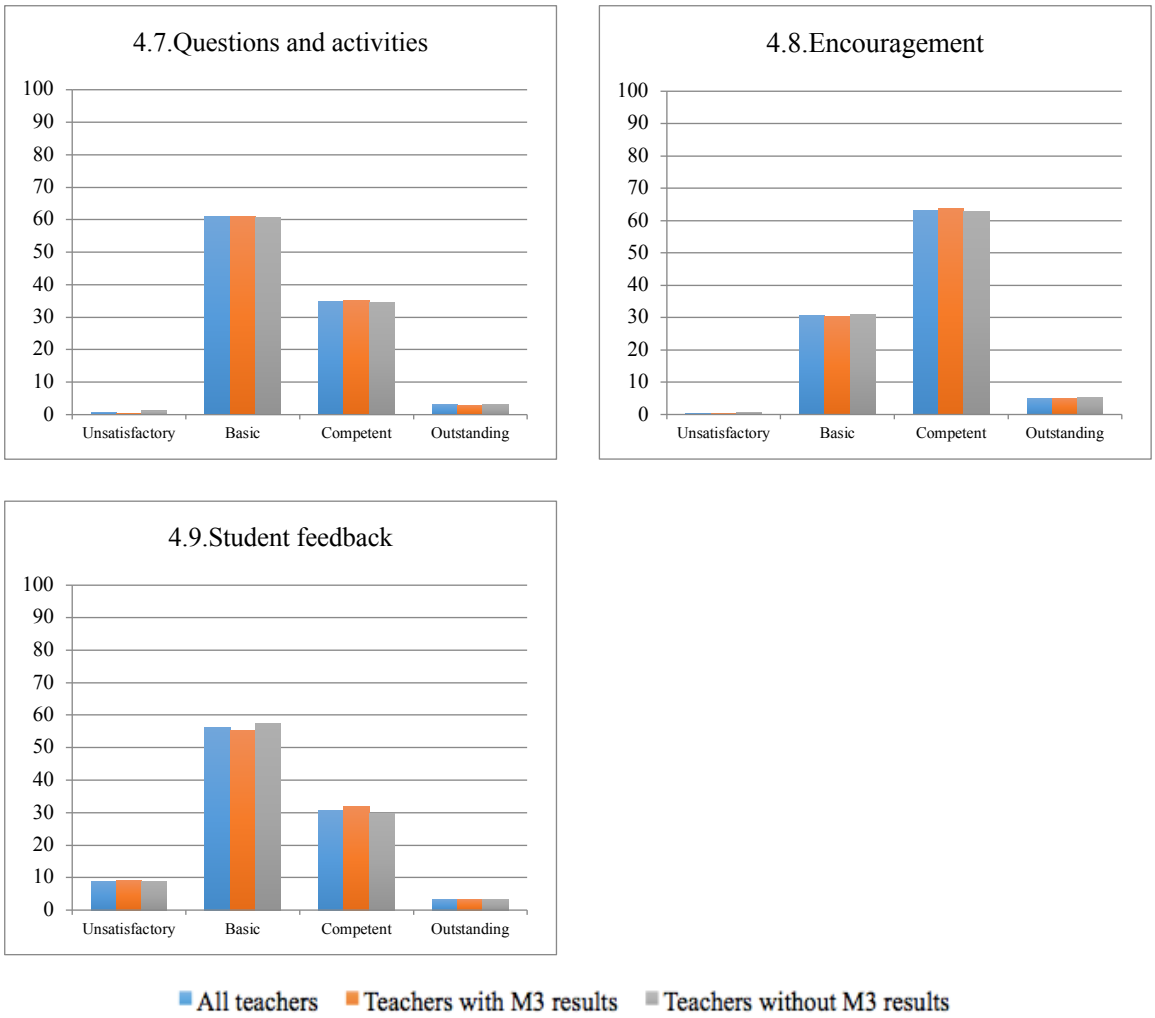


*Figure 4.3.* Percent of responses for the 9 Module 2 portfolio indicators for all teachers, and teachers whose M3 was and was not taken into consideration for the final score.

Evidence for Module 3 was submitted by almost 80% of the teachers evaluated in 2017 ($N$ = 17,953). However, as has been previously mentioned, that evidence was taken into consideration for the teacher's final score when it benefited their results. Significantly, 20% of the teachers did not even turn in evidence to be evaluated in this module. Table 4.3 describes the frequencies and percentages of the four portfolio indicators for Module 3: for the whole sample of teachers, for those teachers whose

Module 3 evaluation was taken into account for their final portfolio score, and those teachers whose Module 3 was not taken into account for their final score but who also submitted evidence to be corrected.

It is possible to observe that for all the teachers evaluated, most were rated as basic or below in all of the four indicators. The indicator with the highest percentage of teachers evaluated as unsatisfactory was the *value of collaborative work for professional development* (unsatisfactory=36%). Unlike the two previously analyzed modules, the distribution pattern of the teachers in each category rate was different between the whole sample and the sample of teachers whose Module 3 was used in their final score. Those teachers who had a score for Module 3 were mainly rated as competent or higher in the four indicators of the module, with the only exception of the indicator 5.4: *reflection on the impact of the collaborative work experience*. Conversely, those teachers whose Module 3 was not considered for their final score were mainly rated as unsatisfactory or basic.

Between both groups of teachers analyzed, significant differences were found in the proportion of teachers in each category for all of the four indicators from Module 3, For all of the indicators, the group of teachers with a Module 3 score showed a higher proportion of teachers as competent or outstanding than the teachers without Module 3 (Ind.5.1: 63.35% vs. 20.45%, $\chi^2_{(3)}$=4,002.89, $p < 0.01$; Ind.5.2: 67.65% vs. 19.55%, $\chi^2_{(3)}$=4,962.76, $p < 0.01$; Ind.5.3: 45.82% vs. 8.89%, $\chi^2_{(3)}$=3,896.76, $p < 0.01$; Ind.5.4: 50.63% vs. 12.48%, $\chi^2_{(3)}$=3,815.47, $p < 0.01$).

Table 4.3.

*Frequencies and percentages for the 4 Module 3 portfolio indicators for all teachers, and teachers whose M3 was and was not taken into consideration for the final score*

| Indicator | All teachers | | Teachers with M3 results | | Teachers without M3 results | |
|---|---|---|---|---|---|---|
| | Frequency | Percent | Frequency | Percent | Frequency | Percent |
| 5.1.Collaborative work suitability | | | *** | | | |
| Unsatisfactory | 1,930 | 10.75 | 303 | 2.70 | 1,627 | 24.15 |
| Basic | 7,539 | 41.99 | 3,807 | 33.94 | 3,732 | 55.40 |
| Competent | 6,756 | 37.63 | 5,478 | 48.84 | 1,278 | 18.97 |
| Outstanding | 1,728 | 9.63 | 1,628 | 14.51 | 100 | 1.48 |
| 5.2.Quality of professional dialogue | | | *** | | | |
| Unsatisfactory | 3,193 | 17.84 | 599 | 5.34 | 2,594 | 38.84 |
| Basic | 5,807 | 32.45 | 3,029 | 27.01 | 2,778 | 41.60 |
| Competent | 7,302 | 40.81 | 6,049 | 53.93 | 1,253 | 18.76 |
| Outstanding | 1,592 | 8.90 | 1,539 | 13.72 | 53 | 0.79 |
| 5.3.Value of collaborative work for professional development | | | *** | | | |
| Unsatisfactory | 6,458 | 36.28 | 2,293 | 20.47 | 4,165 | 63.16 |
| Basic | 5,621 | 31.58 | 3,778 | 33.72 | 1,843 | 27.95 |
| Competent | 5,477 | 30.77 | 4,898 | 43.72 | 579 | 8.78 |
| Outstanding | 242 | 1.36 | 235 | 2.10 | 7 | 0.11 |
| 5.4. Reflection on the impact of the collaborative work experience | | | *** | | | |
| Unsatisfactory | 2,996 | 16.83 | 688 | 6.14 | 2,308 | 35.00 |
| Basic | 8,307 | 46.67 | 4,844 | 49.23 | 3,463 | 52.52 |
| Competent | 5,951 | 33.44 | 5,138 | 45.86 | 813 | 12.33 |
| Outstanding | 544 | 3.06 | 534 | 4.77 | 10 | 0.15 |

*p<0.05; **p<0.01;***p<0.001 (Chi-square test)

Note: Chi-square test with significance levels used for differences between the teachers whose Module 3 was taken into account for the final score, and those teachers whose Module 3 was not included.

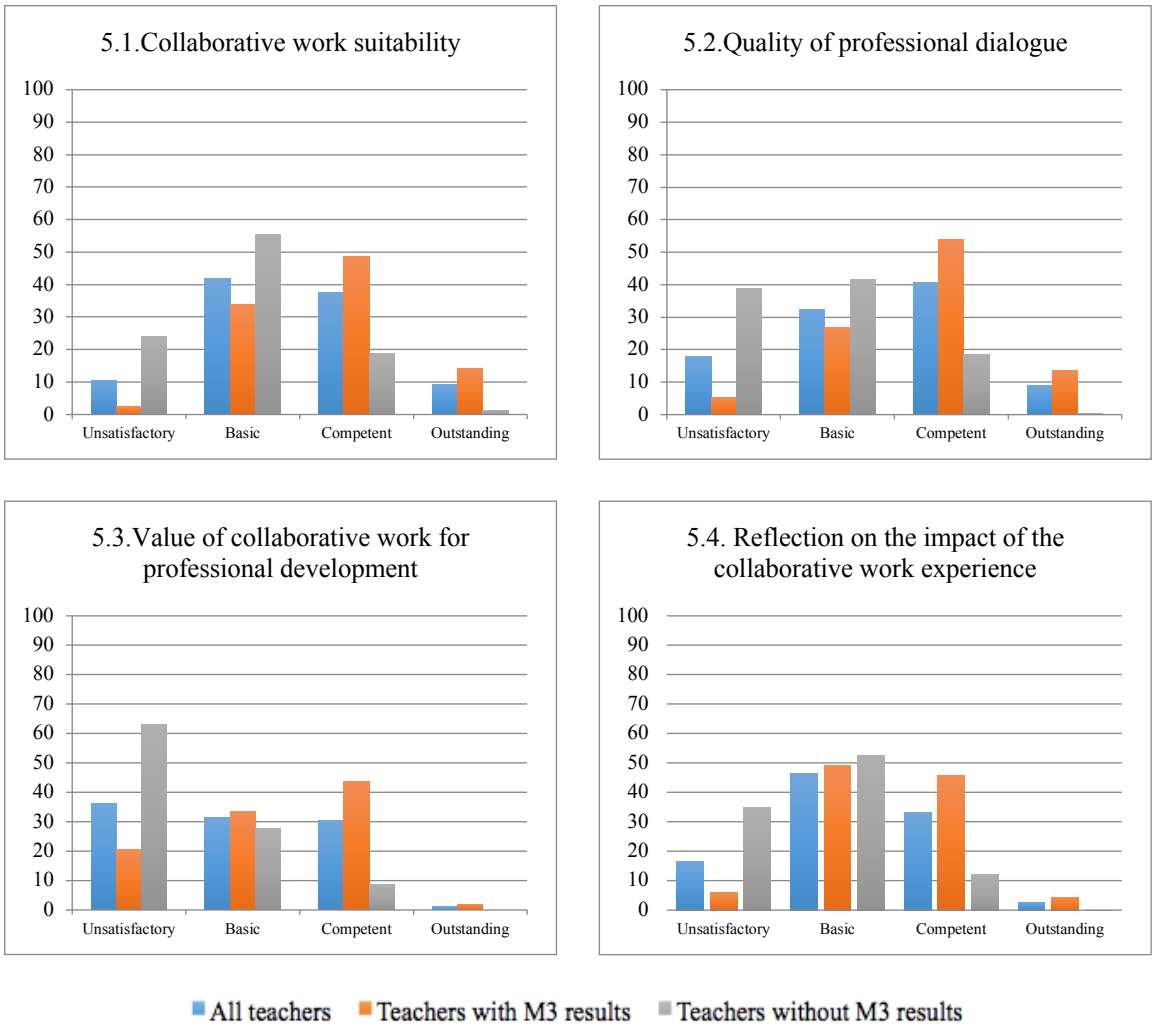


*Figure 4.4.* Percent of responses for the 4 Module 3 portfolio indicators for all teachers, teachers whose M3 was and was not taken into consideration for the final score.

In summary, from the descriptive statistics presented above, it is possible to observe that in general, teachers whose Module 3 was included in their final score were evaluated as competent or outstanding in a higher proportion than those teachers whose Module 3 was not taken into account for the final portfolio score. The difference in proportion in favor of teachers with Module 3 score was significant for the seven indicators of Module 1, and for four of the nine indicators of Module 2. On the other

hand, for Module 2, in only two of the nine indicators the proportion of teachers with a higher proportion as competent and outstanding was seen for those whose Module 3 was not included in their final score. Finally, for Module 3, significant differences were found between both groups of teachers in all of the indicators, with a higher proportion of teachers whose Module 3 was part of their final score in the higher categories. The biggest differences in terms of the proportion between both groups was found in this module, which could mainly be due to the fact that this module is not mandatory. Therefore, from the group of teachers whose Module 3 was not included in their final score, even though some of them submitted evidence to be evaluated in Module 3, that evidence was as not as good.

## Exploratory Confirmatory Factor Analysis (ECFA)

Using the 20 portfolio indicators for the whole sample of teachers evaluated and for those teachers whose Module 3 was taken into consideration for their final score, and using 16 portfolio indicators for those teachers whose Module 3 was not taken into consideration for the final score, five separate exploratory confirmatory factor analysis (ECFA) with oblique Geomin rotation (Asparouhov & Muthén, 2009) models were sequentially tested in each sample (*N*=21,982; *N*=11,216; *N*=10,216, missing=550), representing factor models with one to five dimensions.

The results for each model were evaluated using the three different fit indexes: the comparative fit index (CFI), the root mean square error of approximation (RMSEA), and the standardized root mean square residual (SRMR). The cutoff criteria used for each index was: 1) CFI value close to 0.95 or greater; 2) RMSEA value close to 0.06 or below; and 3) SRMR value close to 0.08 or below (Hu & Bentler, 1995).

For the uni-factor model, the results indicated that the model did not fit the data well ($\chi^2$ = 44,069.5, $p$<.001; RMSEA = .110, CFI = .56, SRMR = .096). The MI results for the 1-factor model suggested that the model could improve with the addition of a covariance between the residual associated with indicator 2.1 *(evaluation used for correction*) and indicator 2.2 (*relationship between assessment and objectives*). The results with the covariate addition improved the uni-factor model ($\chi^2$ = 32,962.8, $p$<.001; RMSEA = .095, CFI = .671, SRMR = .089), however the results still did not fit the data well. In order to improve the model, and taking into consideration the MI results, two other covariances were added one by one (Indicators 5.1 with 5.2 and 5.2 with 5.4), in two successive models for 1-factor, However, the results still did not fit the data well ($\chi^2$

= 29,337.7, $p$<.001; RMSEA = .09, CFI = .708, SRMR = .084; $\chi^2$ = 27,577.4, $p$<.001; RMSEA = .088, CFI = .725, SRMR = .082). Finally, for parsimony no more modifications were included in the 1-factor solution model.

A two-factor model was tested later. Adding one more factor, the model improved (two-factor model: $\chi^2$ = 17,577.3, $p$ <.001; RMSEA = .073, CFI = .825, SRMR = .05), but it still did not fit the data well. In order to improve the model, two covariances were added one by one (Indicators 2.1 with 2.2 and 5.1 with 5.2) in two successive models for the 2-factors solution, considering MI results. However, the results still did not fit the data well ($\chi^2$ = 8,004.9, $p$<.001; RMSEA = .049, CFI = .921, SRMR = .036; $\chi^2$ = 6,064.5, $p$ <.001; RMSEA = .053, CFI = .032). For parsimony, no longer modifications were included in the 2-factor solutions model.

Reasonable fit was detected in the solution with three factors ($\chi^2$ =3,030.9, $p$ <.001; RMSEA = .032, CFI = .971, SRMR = .021), when adding a covariance between the residual associated with indicator 2.1 and 2.2. Good model fit was detected in solutions with four ($\chi^2$ = 2,019.0, $p$ <.001; RMSEA = .028, CFI = .981, SRMR = .017), and five factors ($\chi^2$ = 1,538.2, $p$ <.001; RMSEA = .026, CFI = .986, SRMR = .015), although the addition of those factors did not substantially change RMSEA and CFI.

Different score solutions for the whole sample of teachers evaluated in 2017 are presented in Table 4.4.

Table 4.4.
*Comparison of Exploratory Confirmatory Factor Analysis Models for portfolio indicators*

| **Models** | **$\chi^2$** | **df** | ***p*** | **RMSEA** | **SRMR** | **CFI** | **TLI** |
|---|---|---|---|---|---|---|---|
| 1 factor | 27,577.4 | 167 | >0.001 | .088 | .082 | .725 | .687 |
| 2 factors | 6,064.5 | 149 | >0.001 | .043 | .032 | .941 | .924 |
| **3 factors** | **3,030.9** | **132** | **>0.001** | **.032** | **.021** | **.971** | **.958** |
| 4 factors | 2,019.0 | 115 | >0.001 | .028 | .017 | .981 | .968 |
| 5 factors | 1,538.2 | 100 | >0.001 | .026 | .015 | .986 | .973 |

*Note*. RMSEA= Root mean square error of approximation; SRMR= Standardized root mean square residual; CIF= Comparative Fit Index; TLI=Tucker-Lewis Index

Also, the solutions with three factors obtained had exactly the same structure as the proposed portfolio structure when the loading is larger than 0.3 (with the only exception of the indicator 4.8, with loading of .289). Finally, correlations between factors were generally small (.18~.39), implying that the three dimensions represented distinct aspects of the portfolio used to evaluate Chilean teachers. Table 4.5 shows the results.

Table 4.5.
*Exploratory Confirmatory Factor Analysis of the Portfolio Indicators from Teacher Evaluation 2017 (N=21,432)*

| Indicators | Factor | | | Dimension |
|---|---|---|---|---|
| | 1 | 2 | 3 | |
| 1.1.Formation of learning objectives | **0.315** | 0.014 | 0.015 | |
| 1.2.Relationship between activities and objectives | **0.463** | -0.044 | -0.008 | |
| 2.1. Evaluation and rubrics used for correction | **0.415** | 0.021 | -0.046 | |
| 2.2.Relationship between assessment and objectives | **0.504** | -0.014 | -0.037 | **Pedagogical materials** |
| 2.3.Analysis and use of assessment results | **0.549** | 0.081 | -0.002 | |
| 3.1.Analyses based on students' characteristics | **0.464** | 0.186 | 0.018 | |
| 3.2.Use of error for learning | **0.448** | 0.200 | 0.015 | |
| 4.1.Class environment | 0.022 | -0.056 | **0.522** | |
| 4.2.Quality of the start of class | 0.015 | 0.005 | **0.590** | |
| 4.3.Quality of the end of class | 0.008 | 0.010 | **0.621** | |
| 4.4.Contribution of the activities to the achievement of the class objectives | 0.186 | -0.026 | **0.360** | |
| 4.5.Curricular emphasis on the subject | -0.036 | 0.038 | **0.391** | **Video recording of a class** |
| 4.6.Clear explanations | -0.019 | -0.014 | **0.661** | |
| 4.7.Questions and activities | -0.064 | 0.015 | **0.730** | |
| 4.8.Encouragement | 0.095 | -0.012 | ***0.289*** | |
| 4.9.Student feedback | 0.020 | 0.016 | **0.574** | |
| 5.1.Collaborative work suitability | -0.002 | **0.678** | 0.014 | |
| 5.2.Quality of professional dialogue | -0.009 | **0.780** | 0.000 | |
| 5.3.Value of collaborative work for professional development | 0.198 | **0.489** | -0.016 | **Collaborative work** |
| 5.4. Reflection on the impact of the collaborative work experience | 0.146 | **0.571** | 0.000 | |

*Notes.* Extraction method; maximum likelihood; Rotation method; Oblimin with Kaiser normalization. Loadings larger than .30 are in bold.
Model fit: $\chi^2$ = 3,030.9, *df* = 132, $p < .001$; RMSEA = .032; CFI = .971; TLI = .958.

Given the aforementioned model outcomes for the whole sample of teachers evaluated, among all models tested, the three-factor solution offered the clearest factor structure in accordance with our theoretical expectations, representing three underlying dimensions evaluated by the portfolio. The first dimension, *pedagogical material*, consisted of the first seven portfolio indicators (loadings = .32~.55) capturing planning, assessment, and reflection. The second dimension, *video recorded class*, included the nine indicators of portfolio Module 2 (loadings = .29~.73) measuring the quality of a real class given by the teacher. The third dimension, *collaborative work*, represented the teacher's work and cooperation with their colleagues (loadings = .49~.78). All the indicators loaded mainly onto one dimension.

The results presented above confirmed that the hypothesized tree-factor structure of the portfolio was replicated in this analysis. ECFA results confirmed how well the analyzed indicators represent a smaller number of constructs, since all the indicators significantly loaded in their respective subscale, and according to the CFRA, RMSEA, SRMR indices the three dimension structure fits the data well.

For the sample of teachers whose Module 3 was not part of their final score results, different score solutions are presented in Table 4.6.

Table 4.6.
*Comparison of Exploratory Confirmatory Factor Analysis Models for portfolio indicators*

| **Models** | $\chi^2$ | **df** | ***p*** | **RMSEA** | **SRMR** | **CFI** | **TLI** |
|---|---|---|---|---|---|---|---|
| 1 factor | 7,161.5 | 102 | >0.001 | .082 | .067 | .814 | .782 |
| **2 factors** | **935.4** | **87** | **>0.001** | **.031** | **.021** | **.978** | **.969** |
| 3 factors | 722.7 | 74 | >0.001 | .029 | .018 | .983 | .972 |
| 4 factors | 328.5 | 61 | >0.001 | .021 | .012 | .993 | .986 |
| 5 factors | 270.6 | 50 | >0.001 | .021 | .011 | .994 | .986 |

*Note*. RMSEA= Root mean square error of approximation; SRMR= Standardised root mean square residual; CIF= Comparative Fit Index; TLI=Tucker-Lewis Index

The results indicated that a uni-factor model did not fit the data well ($\chi^2$ = 7,161.5, $p$ <.001; RMSEA = .082, CFI = .814, SRMR = .067), even when two covariance between the residuals of variables were added (Indicator 2.1 and 2.2, and 3.1 with 3.2,). For a two-factor solution, the model improved when adding a covariance between the residual associated with indicator 2.1 and 2.2, and 3.1 with 3.2. Good model fit was detected (two-factor model: $\chi^2$ = 935.4, $p$ <.001; RMSEA = .031, CFI = .978, SRMR = .021). In addition, good model fit was detected in solutions with three ($\chi^2$ = 722.7, $p$<.001; RMSEA = .029, CFI = .983, SRMR = .018), four ($\chi^2$ = 328.5, $p$ <.001; RMSEA = .021, CFI = .993, SRMR = .012), and five factors ($\chi^2$ = 270.6, $p$ <.001; RMSEA = .021, CFI = .994, SRMR = .021), although the addition of those factors did not substantially change RMSEA and CFI. Also, the solution for more than two factors showed the two first indicators with loading lower than 0.3 in any of the factors. Additionally, the solutions with two factors had exactly the same structure as the proposed portfolio structure. Finally, correlation between factors was small (.32), implying that the two dimensions represented distinct aspects of the portfolio used to evaluate Chilean teachers. Results are presented in Table 4.7.

Table 4.7.
*Exploratory Confirmatory Factor Analysis of the portfolio indicators from Teacher Evaluation 2017 (N=10,216)*

| Indicators | Factor | | Dimension |
|---|---|---|---|
| | 1 | 2 | |
| 1.1.Formation of learning objectives | **0.356** | 0.058 | |
| 1.2.Relationship between activities and objectives | **0.473** | 0.006 | |
| 2.1. Evaluation and rubrics used for correction | **0.460** | -0.033 | |
| 2.2.Relationship between assessment and objectives | **0.538** | -0.031 | **Pedagogical materials** |
| 2.3.Analysis and use of assessment results | **0.623** | -0.004 | |
| 3.1.Analyses based on students' characteristics | **0.491** | 0.057 | |
| 3.2.Use of error for learning | **0.457** | 0.053 | |
| 4.1.Class environment | 0.020 | **0.555** | |
| 4.2.Quality of the start of class | 0.027 | **0.593** | |
| 4.3.Quality of the end of class | 0.000 | **0.624** | |
| 4.4.Contribution of the activities to the achievement of the class objectives | 0.154 | **0.387** | **Video recording of a class** |
| 4.5.Curricular emphasis on the subject | 0.002 | **0.391** | |
| 4.6.Clear explanations | -0.018 | **0.668** | |
| 4.7.Questions and activities | -0.072 | **0.753** | |
| 4.8.Encouragement | 0.102 | **0.303** | |
| 4.9.Student feedback | 0.005 | **0.594** | |

*Notes.* Extraction method; Weighted least square mean and variance adjusted (WLSMV); Rotation method; Geomin with Kaiser normalization. Loadings larger than .30 are in bold
Model fit: $\chi^2 = 935.4$, df = 87, $p < .001$; RMSEA = .031; CFI = .978; TLI = .969.

Looking at the model outcomes for the sample of teachers whose Module 3 was not considered for their final score, among all models tested, the two-factor solution offered the clearest factor structure in accordance with our theoretical expectations. The two-factor solution represented two underlying dimensions evaluated by the portfolio:

*pedagogical material*, encompassing the first seven portfolio indicators (loadings = .36~.62), and *video recorded class*, made up of the nine indicators of portfolio Module 2 (loadings = .30~.75). All of the indicators loaded mainly onto one dimension.

For the teachers whose Module 3 was part of their final score results, different factor solutions are presented in Table 4.8.

Table 4.8.
*Comparison of Exploratory Confirmatory Factor Analysis Models for portfolio indicators*

| **Models** | **$\chi^2$** | **df** | ***p*** | **RMSEA** | **SRMR** | **CFI** | **TLI** |
|---|---|---|---|---|---|---|---|
| 1 factor | 10,405.4 | 168 | >0.001 | .074 | .068 | .754 | .722 |
| **2 factors** | **1,609.6** | **149** | **>0.001** | **.030** | **.023** | **.965** | **.955** |
| **3 factors** | **1,269.6** | **132** | **>0.001** | **.028** | **.020** | **.973** | **.961** |
| 4 factors | 1,050.0 | 115 | >0.001 | .027 | .018 | .978 | .963 |
| 5 factors | 682.5 | 100 | >0.001 | .023 | .015 | .986 | .973 |

*Note*. RMSEA= Root mean square error of approximation; SRMR= Standardised root mean square residual; CIF= Comparative Fit Index; TLI=Tucker-Lewis Index

The results indicated that a uni-factor model did not fit the data well, even when two covariance between the residuals of variables were added (Indicator 2.1 with 2.2, and 5.1 with 5.2; $\chi^2$ = 10,405.4, $p$ <.001; RMSEA = .074, CFI = .754, SRMR = .068). Adding one more factor and two covariance between variables (2.1 with 2.2, and 5.1 with 5.2), the model improved (two-factor model: $\chi^2$ = 1,609.6, $p$ <.001; RMSEA = .030, CFI = .965, SRMR = .023), indicating a good model fit. Good model fit was also detected in solutions with three ($\chi^2$ = 1,269.6, $p$ <.001; RMSEA = .028, CFI = .973, SRMR = .020), four ($\chi^2$ = 1,050.0, $p$ <.001; RMSEA = .027, CFI = .978, SRMR = .018), and five factors

($\chi^2$ = 682.5, $p$ <.001; RMSEA = .023, CFI = .986, SRMR = .015), although the addition of those factors did not substantially change RMSEA and CFI.

The solution for two factors showed that indicators from Module 1 and Module 3 loaded into the same factor. Indicators from Module 2 loaded into a second factor. The solution with two factors is presented in Table 4.9.

Table 4.9.
*Exploratory Confirmatory Factor Analysis of the portfolio indicators from Teacher Evaluation 2017 (N=11,216)*

| **Indicators** | **Factor** | | **Dimension** |
|---|---|---|---|
| | **1** | **2** | |
| 1.1.Formation of learning objectives | ***0.284*** | -0.034 | |
| 1.2.Relationship between activities and objectives | **0.386** | -0.034 | |
| 2.1. Evaluation and rubrics used for correction | **0.392** | -0.072 | **Pedagogical materials and collaborative work** |
| 2.2.Relationship between assessment and objectives | **0.432** | -0.055 | |
| 2.3.Analysis and use of assessment results | **0.538** | -0.010 | |
| 3.1.Analyses based on students' characteristics | **0.539** | 0.001 | |
| 3.2.Use of error for learning | **0.530** | 0.008 | |
| 4.1.Class environment | -0.021 | **0.479** | |
| 4.2.Quality of the start of class | 0.009 | **0.588** | |
| 4.3.Quality of the end of class | 0.024 | **0.620** | |
| 4.4.Contribution of the activities to the achievement of the class objectives | 0.157 | **0.333** | **Video recording of a class** |
| 4.5.Curricular emphasis on the subject | -0.001 | **0.385** | |
| 4.6.Clear explanations | -0.011 | **0.649** | |
| 4.7.Questions and activities | -0.019 | **0.703** | |
| 4.8.Encouragement | 0.084 | ***0.269*** | |
| 4.9.Student feedback | 0.058 | **0.552** | |
| 5.1.Collaborative work suitability | ***0.269*** | 0.069 | **Pedagogical materials** |

| | | | |
|---|---|---|---|
| 5.2.Quality of professional dialogue | **0.365** | 0.041 | **and collaborative** |
| 5.3.Value of collaborative work for professional development | **0.454** | 0.026 | **work** |
| 5.4. Reflection on the impact of the collaborative work experience | **0.472** | 0.031 | |

*Notes.* Extraction method; Weighted least square mean and variance adjusted (WLSMV); Rotation method; Geomin with Kaiser normalization. Loadings larger than .30 are in bold and loadings larger than .20 are in italic and bold.
Model fit: $\chi^2$ = 1,609.6, df = 149, $p < .001$; RMSEA = .030; CFI = .965; TLI = .955.

The results for the sample of teachers whose Module 3 was not part of their final score confirmed that the hypothesized two-factor structure of the portfolio was replicated in this analysis. ECFA results confirmed how well the analyzed indicators represent a smaller number of constructs, since all the indicators significantly loaded in their respective subscale, and according to the CFRA, RMSEA, SRMR indices the two dimension structure fits the data well.

Conversely, looking at the model outcomes for the sample of teachers whose Module 3 was considered for their final score, among all models tested, both the two and the three-factor solution did not offer the clearest factor structure in accordance with our theoretical expectations. For the two-factor solution, Module 3 indicators loaded into the same factor as the indicators of Module 1, which could indicate one big factor that groups the pedagogical material and the collaborative work. For the three-factor solution, two of the Module 1 indicators split into a single factor related to assessment. Also, two of the Module 1 indicators showed loading lower than .25. The other indicators from Module 1 loaded into the same factor as the four Module 3 indicators. For both solutions, all the

indicators from Module 2 loaded mainly onto one single factor. Table 4.10 shows the solution with three factors.

These results did not confirm that the theoretical three-factor structure of the portfolio was replicated in the ECFA analysis. The indicators did not significantly load in their respective subscales. Thus, the theoretical three-factor structure when teachers whose M3 results were not included in their final score was removed, was not supported by the empirical information.

Table 4.10.
*Exploratory Confirmatory Factor Analysis of the portfolio indicators from Teacher Evaluation 2017 (N=11,216)*

| Indicators | Factor | | | Dimension |
|---|---|---|---|---|
| | 1 | 2 | 3 | |
| 1.1.Formation of learning objectiv | 0.082 | ***0.226*** | -0.021 | **Planning, reflection and collaborative work** |
| 1.2.Relationship between activities and objectives | 0.187 | ***0.240*** | 0.005 | |
| 2.1. Evaluation and rubrics used for correction | **0.624** | 0.094 | -0.024 | **Assessment** |
| 2.2.Relationship between assessment and objectives | **0.915** | -0.013 | 0.022 | |
| 2.3.Analysis and use of assessment results | 0.116 | **0.456** | 0.007 | **Planning, reflection and collaborative work** |
| 3.1.Analyses based on students' characteristics | -0.015 | **0.581** | -0.021 | |
| 3.2.Use of error for learning | -0.008 | **0.560** | -0.010 | |
| 4.1.Class environment | 0.020 | -0.050 | **0.488** | **Video recording of a class** |
| 4.2.Quality of the start of class | -0.005 | -0.001 | **0.591** | |
| 4.3.Quality of the end of class | 0.018 | -0.007 | **0.630** | |
| 4.4.Contribution of the activities to the achievement of the class objectives | 0.121 | 0.053 | **0.362** | |
| 4.5.Curricular emphasis on the subject | -0.026 | 0.017 | **0.379** | |
| 4.6.Clear explanations | -0.032 | -0.002 | **0.647** | |

| | | | | |
|---|---|---|---|---|
| 4.7.Questions and activities | -0.027 | 0.014 | **0.702** | |
| 4.8.Encouragement | 0.033 | 0.056 | ***0.275*** | |
| 4.9.Student feedback | 0.005 | 0.043 | **0.556** | |
| 5.1.Collaborative work suitability | 0.023 | ***0.259*** | 0.067 | |
| 5.2.Quality of professional dialogue | 0.034 | **0.350** | 0.040 | **Planning, reflection and collaborative work** |
| 5.3.Value of collaborative work for professional development | 0.011 | **0.464** | 0.015 | |
| 5.4. Reflection on the impact of the collaborative work experience | -0.033 | **0.521** | 0.008 | |

*Notes.* Extraction method; maximum likelihood; Rotation method; Oblimin with Kaiser normalization. Loadings larger than .30 are in bold and loadings larger than .20 are in italic and bold.

Model fit: $\chi^2$ = 1,269.6, df = 132, $p$ < .001; RMSEA = .028; CFI = .973; TLI = .961.

Aim 2: Determine if the portfolio factor structures are invariant across subgroups

For this second aim, I looked for evidence of measurement invariance in order to determine if the portfolio works equivalently across different populations such as school location (e.g., teachers from urban schools and teachers from rural schools), or teaching levels (e.g., preschool teachers and high school teachers).

Descriptive statistics related to the portfolio indicator score for each one of the groups of teachers who were included in the analysis of invariance are described in the following section.

Descriptive statistics for teacher school location.

Table 4.11 describes the frequencies and percentages of rural ($N$ = 4,851) and urban ($N$ = 16,577; missing = 550) teachers' performance level rating for each one of the seven portfolio indicators for Module 1. The distribution of teachers depending on their performance level was the same one that was described for the whole sample of teachers in aim 1. With respect to the differences between rural and urban teachers, significant differences were found in the proportion of teachers rated in each category for both groups in four of the seven indicators from Module 1. In two of those indicators rural teachers scored proportionally higher as competent or outstanding (Ind.1.2: 62.95% vs. 65.68%, $\chi^2_{(3)}$=13.4, $p < 0.01$; Ind.2.1: 37.91% vs. 38.85%, $\chi^2_{(3)}$=22.03, $p < 0.01$). For the other two indicators with significant differences, urban teachers' results showed a higher proportion as competent or better (Ind.2.2: 48.94% vs. 48.05%, $\chi^2_{(3)}$=13.6, $p < 0.01$; Ind.3.2: 65.65% vs. 62.92%, $\chi^2_{(3)}$=14.1, $p < 0.01$).

Table 4.11.

*Frequencies and percentages for the 7 Module 1 portfolio indicators for teachers based on school location*

| Indicator | Urban teachers | | Rural teachers | |
|---|---|---|---|---|
| | **Frequency** | **Percent** | **Frequency** | **Percent** |
| 1.1.Formation of learning objectives | | | | |
| Unsatisfactory | 373 | 2.25 | 95 | 1.96 |
| Basic | 1,271 | 7.67 | 365 | 7.52 |
| Competent | 12,677 | 76.47 | 3,713 | 76.54 |
| Outstanding | 2,256 | 13.61 | 678 | 13.98 |
| 1.2.Relationship between activities and objectives | | | ** | |
| Unsatisfactory | 2,401 | 14.48 | 626 | 12.90 |
| Basic | 3,741 | 22.57 | 1,039 | 21.42 |
| Competent | 10,066 | 60.72 | 3,078 | 63.45 |
| Outstanding | 369 | 2.23 | 108 | 2.23 |
| 2.1. Evaluation and rubrics used for correction | | | *** | |
| Unsatisfactory | 2,828 | 17.19 | 692 | 14.37 |
| Basic | 7,358 | 44.90 | 2,253 | 46.78 |
| Competent | 6,022 | 36.74 | 1,818 | 37.75 |
| Outstanding | 191 | 1.17 | 53 | 1.10 |
| 2.2.Relationship between assessment and objectives | ** | | | |
| Unsatisfactory | 2,836 | 17.30 | 755 | 15.68 |
| Basic | 5,534 | 33.76 | 1,747 | 36.27 |
| Competent | 7,017 | 42.81 | 2,013 | 41.80 |
| Outstanding | 1,004 | 6.13 | 301 | 6.25 |
| 2.3.Analysis and use of assessment results | | | | |
| Unsatisfactory | 3,079 | 18.84 | 863 | 17.98 |
| Basic | 5,733 | 35.07 | 1,735 | 36.15 |
| Competent | 7,087 | 43.35 | 2,081 | 43.35 |
| Outstanding | 448 | 2.74 | 121 | 2.52 |
| 3.1.Analyses based on students' characteristics | | | | |
| Unsatisfactory | 1,706 | 10.33 | 493 | 10.19 |
| Basic | 7,255 | 43.95 | 2,199 | 45.43 |
| Competent | 7,171 | 43.44 | 2,046 | 42.27 |
| Outstanding | 376 | 2.28 | 102 | 2.11 |
| 3.2.Use of error for learning | ** | | | |
| Unsatisfactory | 1,549 | 9.41 | 518 | 10.72 |
| Basic | 4,108 | 24.95 | 1,273 | 26.36 |
| Competent | 10,360 | 62.93 | 2,920 | 60.46 |
| Outstanding | 447 | 2.72 | 119 | 2.46 |

*p<0.05; **p<0.01;***p<0.001 (Chi-square test)

Note: Chi-square test with significance levels used for differences between the teachers from urban schools and teachers from rural schools.

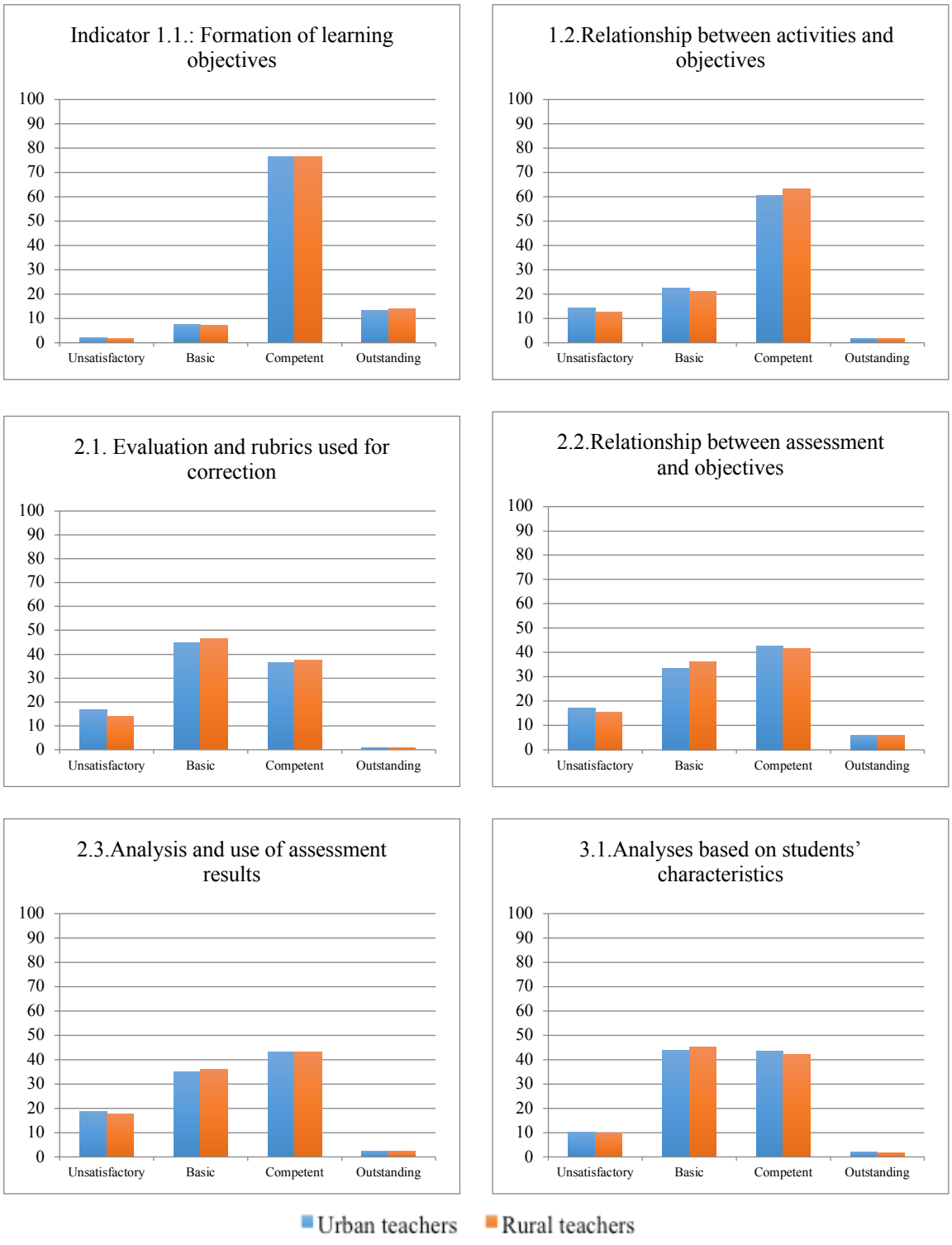


*Figure 4.5.* Percent of responses for the 7 Module 1 portfolio indicators for teachers based on school location.

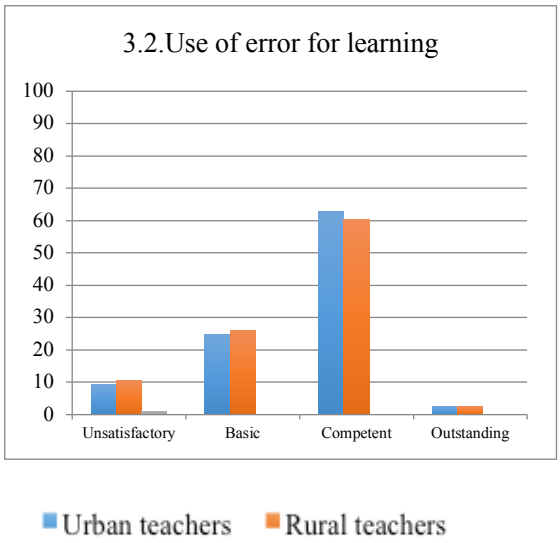


*Figure 4.6.* Percent of responses for the 7 Module 1 portfolio indicators for teachers based on school location.

Frequencies and percentages of rural and urban teachers' performance level rating for the nine portfolio indicators of Module 2 are presented in Table 4.12. The distribution of teachers within performance levels for each indicator of Module 2 is the same one that was previously described for the whole sample of teachers. Significant differences in the proportion of score classification of rural and urban teachers were found for all of the nine indicators of Module 2. For eight of the nine indicators, the group of urban teachers was rated at a higher proportion as competent or outstanding than the rural teachers (Ind.4.2: 69.49% vs. 67.38%, $\chi^2_{(3)}=9.01$, $p = 0.03$; Ind.4.3: 48.03% vs. 45.57%, $\chi^2_{(3)}=28.67$, $p < 0.01$; Ind.4.4: 74.34% vs. 70.49%, $\chi^2_{(3)}=62.49$, $p < 0.01$; Ind.4.5: 24.32% vs.20.91%, $\chi^2_{(3)}=39.46$, $p < 0.01$; Ind.4.6: 36.48% vs. 31.55%, $\chi^2_{(3)}=48.8$, $p < 0.01$; Ind.4.7: 39.16% vs. 34.79%, $\chi^2_{(3)}=59.4$, $p < 0.01$; Ind.4.8: 69.42% vs. 65.5%, $\chi^2_{(3)}=27.47$, $p < 0.01$; Ind.4.9: 34.7% vs. 33.33%, $\chi^2_{(3)}=11.41$, $p = 0.01$). The only

exception was the indicator that measures class environment, in which rural teachers were evaluated at a higher proportion as competent or better (Ind.4.1: 93.56% vs. 95.34%, $\chi^2_{(3)} = 45.99$, $p < 0.01$).

Table 4.12.

*Frequencies and percentages for the 9 Module 2 portfolio indicators for teachers based on school location*

| Indicator | Urban teachers | | Rural teachers | |
|---|---|---|---|---|
| | Frequency | Percent | Frequency | Percent |
| 4.1.Class environment | | | *** | |
| Unsatisfactory | 159 | 0.96 | 53 | 1.09 |
| Basic | 905 | 5.46 | 173 | 3.57 |
| Competent | 14,697 | 88.68 | 4,450 | 91.77 |
| Outstanding | 813 | 4.91 | 173 | 3.57 |
| 4.2.Quality of the start of class | * | | | |
| Unsatisfactory | 744 | 4.49 | 219 | 4.52 |
| Basic | 4,312 | 26.02 | 1,363 | 28.11 |
| Competent | 11,291 | 68.12 | 3,208 | 66.16 |
| Outstanding | 227 | 1.37 | 59 | 1.22 |
| 4.3.Quality of the end of class | *** | | | |
| Unsatisfactory | 1,300 | 7.84 | 380 | 7.84 |
| Basic | 7,314 | 44.13 | 2,259 | 46.59 |
| Competent | 7,382 | 44.54 | 2,108 | 43.47 |
| Outstanding | 578 | 3.49 | 102 | 2.10 |
| 4.4.Contribution of the activities to the achievement of the class objectives | *** | | | |
| Unsatisfactory | 1,410 | 8.51 | 590 | 12.17 |
| Basic | 2,842 | 17.15 | 841 | 17.34 |
| Competent | 12,032 | 72.61 | 3,349 | 69.07 |
| Outstanding | 286 | 1.73 | 69 | 1.42 |
| 4.5.Curricular emphasis on the subject | *** | | | |
| Unsatisfactory | 4,722 | 28.49 | 1,527 | 31.49 |
| Basic | 7,821 | 47.19 | 2,308 | 47.60 |
| Competent | 3,862 | 23.30 | 993 | 20.48 |
| Outstanding | 169 | 1.02 | 21 | 0.43 |
| 4.6.Clear explanations | *** | | | |
| Unsatisfactory | 1,398 | 8.43 | 452 | 9.32 |
| Basic | 9,130 | 55.09 | 2,867 | 59.13 |
| Competent | 5,902 | 35.61 | 1,513 | 31.20 |
| Outstanding | 133 | 0.87 | 17 | 0.35 |
| 4.7.Questions and activities | *** | | | |
| Unsatisfactory | 160 | 0.97 | 50 | 1.03 |
| Basic | 9,924 | 59.88 | 3,112 | 64.18 |
| Competent | 5,874 | 35.44 | 1,599 | 32.98 |
| Outstanding | 616 | 3.72 | 88 | 1.81 |
| 4.8.Encouragement | *** | | | |
| Unsatisfactory | 117 | 0.71 | 42 | 0.87 |
| Basic | 4,951 | 29.87 | 1,631 | 33.64 |
| Competent | 10,602 | 63.97 | 2,938 | 60.59 |
| Outstanding | 904 | 5.45 | 238 | 4.91 |
| 4.9.Student feedback | * | | | |
| Unsatisfactory | 1,472 | 8.88 | 501 | 10.33 |
| Basic | 9,351 | 56.42 | 2,732 | 56.34 |
| Competent | 5,163 | 31.15 | 1,463 | 30.17 |
| Outstanding | 588 | 3.55 | 153 | 3.16 |

*p<0.05; **p<0.01;***p<0.001 (Chi-square test)

Note: Chi-square test with significance levels used for differences between the teachers from urban schools and teachers from rural schools.

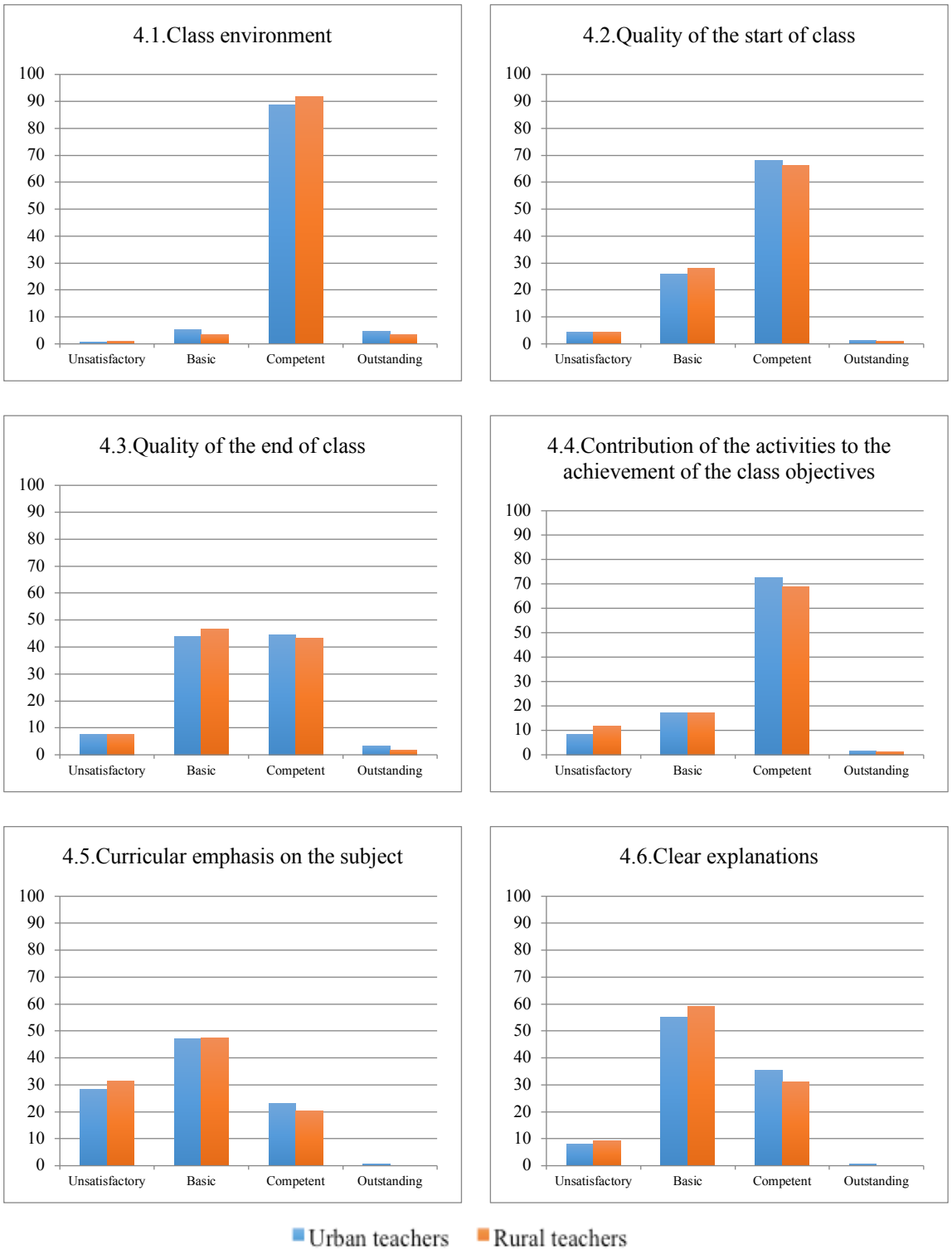


*Figure 4.7.* Percent of responses for the 9 Module 2 portfolio indicators for teachers depending on the school location.

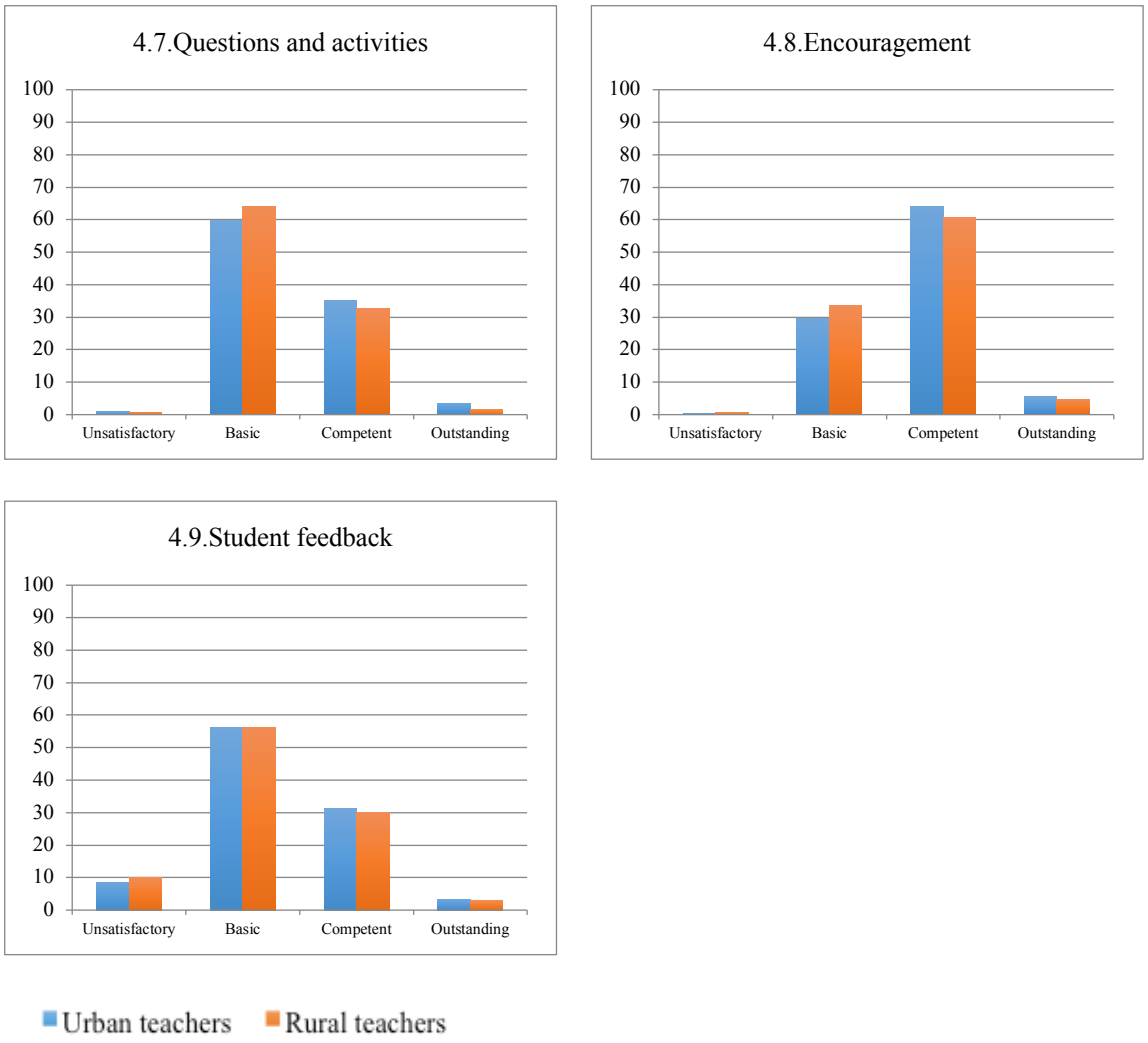


*Figure 4.8.* Percent of responses for the 9 Module 2 portfolio indicators for teachers depending on the school location.

Table 4.13 describes the frequencies and percentages for rural and urban teachers' performance level within the four portfolio indicators of Module 3. The distribution of teachers in each one of the four performance levels was similar to the one that was described for the whole sample of teachers in the results of aim 1. Significant differences between both groups were found in the four indicators; the urban teachers had a higher

proportion of teachers rated as competent or outstanding compared to the rural teachers (Ind.5.1: 48.48% vs. 43.13%, $\chi^2_{(3)}=36.71$, $p < 0.01$; Ind.5.2: 50.3% vs. 47.69%, $\chi^2_{(3)}=10.48$, $p = 0.02$; Ind.5.3: 33.06% vs. 29.02%, $\chi^2_{(3)}=27.55$, $p < 0.01$; Ind.5.4: 36.99% vs. 34.8%, $\chi^2_{(3)}=12.47$, $p < 0.01$).

Table 4.13.
*Frequencies and percentages for the 4 Module 3 portfolio indicators for teachers based on school location*

| Indicator | Urban teachers | | Rural teachers | |
|---|---|---|---|---|
| | Frequency | Percent | Frequency | Percent |
| 5.1.Collaborative work suitability | *** | | | |
| Unsatisfactory | 1,444 | 10.42 | 486 | 11.86 |
| Basic | 5,694 | 41.10 | 1,845 | 45.01 |
| Competent | 5,346 | 38.59 | 1,410 | 34.40 |
| Outstanding | 1,370 | 9.89 | 358 | 8.73 |
| 5.2.Quality of professional dialogue | * | | | |
| Unsatisfactory | 2,427 | 17.58 | 766 | 18.74 |
| Basic | 4,435 | 32.12 | 1,372 | 33.57 |
| Competent | 5,682 | 41.15 | 1,620 | 39.64 |
| Outstanding | 1,263 | 9.15 | 329 | 8.05 |
| 5.3.Value of collaborative work for professional development | ** | | | |
| Unsatisfactory | 4,877 | 35.51 | 1,581 | 38.91 |
| Basic | 4,318 | 31.44 | 1,303 | 32.07 |
| Competent | 4,355 | 31.71 | 1,122 | 27.62 |
| Outstanding | 185 | 1.35 | 57 | 1.40 |
| 5.4. Reflection on the impact of the collaborative work experience | ** | | | |
| Unsatisfactory | 2,251 | 16.39 | 745 | 18.34 |
| Basic | 6,403 | 46.62 | 1,904 | 46.86 |
| Competent | 4,645 | 33.82 | 1,306 | 32.14 |
| Outstanding | 436 | 3.17 | 108 | 2.66 |

*p<0.05; **p<0.01;***p<0.001 (Chi-square test)
Note: Chi-square test with significance levels used for differences between the teachers from urban schools and teachers from rural schools.

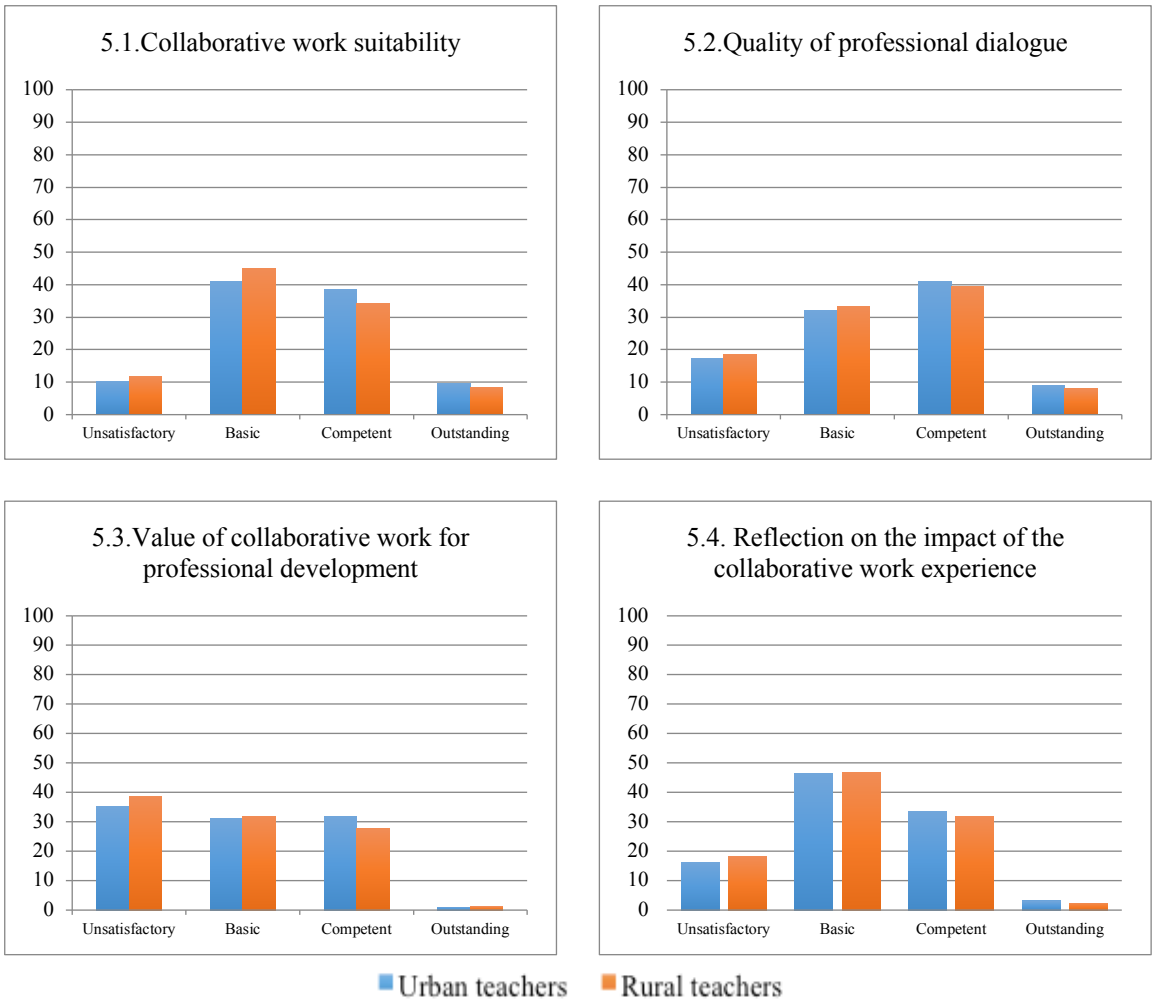


*Figure 4.9.* Percent of responses for the 4 Module 3 portfolio indicators for teachers based on school location.

In summary, from the descriptive statistics of the teache results by school setting (rural/ urban), the results indicated that in general, teachers form urban schools were evaluated as competent or outstanding in a higher proportion than those teachers from rural schools. However, for Module 1, the significant differences between both groups were found in four of the seven indicators; two of them had a higher proportion of teachers with competent and outstanding out of the urban teachers group and the other two did not show a higher proportion for urban teachers. For Module 2, better results of urban teachers was clearer, where eight of the nine indicators showed significant differences between both groups, with a higher proportion of teachers from urban schools evaluated as competent or outstanding. For this module, the only exception was the class environment indicator, in which a higher proportion of rural teachers were evaluated at the highest levels. Finally, for Module 3, significant differences for both groups were found for the four indicators, with a higher proportion of urban teachers evaluated as competent or outstanding compared to the rural teachers.

## Measurement invariance for teacher school location

Measurement invariance for teacher school location was tested within the framework of multigroup confirmatory factor analysis (CFA), using the procedures outlined by Byrne (2012), and conducted using M*plus* 8. A robust weighted least square estimator (WLSMV) was used, which is the recommended methodology to use for categorical ordered data.

Testing for factorial equivalence encompasses a series of nested models, which were progressively evaluated using the chi-square differences DIFFTEST, which

indicates when a relevant deviation for the data from the model is significant. Since chi-square can be affected by the large sample size, which is the case for the present research, other fit measures were used. Thus, RMSEA, CFI, and SRMR fit indexes changes were also used to evaluate the progressive factorial invariance. Cutoff points of change equal to or lower than -.010 in CFI, supplemented by a change equal to or higher than .015 in RMSEA, or a change equal to or higher than .030 (or .010) in SRMR would indicate noninvariance (Chen, 2007).

The Hypothesized Model.

A necessary requisite in testing for multigroup invariances is the establishment of a well-fitting baseline model structure for each group. Once these models are established, they represent the hypothesized multigroup model under test. The model under test in this initial multigroup measurement of variance was the same one that proposed a three-factor portfolio structure solution, with a covariance between indicators 2.1 and 2.2, tested in aim 1 for the whole sample of teachers. It serves as the initial model tested as the establishment of baseline models for the two groups of teachers.

***Establishing Baseline Models***. For rural and urban teachers, CFA was conducted evaluating a three-factor structure of the portfolio with one residual covariance between indicators 2.1 and 2.2, in each sample separately. Model goodness-of-fit statistics for each group of teachers was as follows:

Urban teachers: $\chi^2_{(166)}$ =2,904.6, $p$ <.001; RMSEA = .032, CFI = .964.

Rural teachers: $\chi^2_{(166)}$ = 926.7, $p$ <.001; RMSEA = .031, CFI = .967.

Figure 4.7 shows the representation of the baseline models for both urban and rural teachers, and it provides the foundation with which I tested the series of stringent between-group constraints related to portfolio structure.

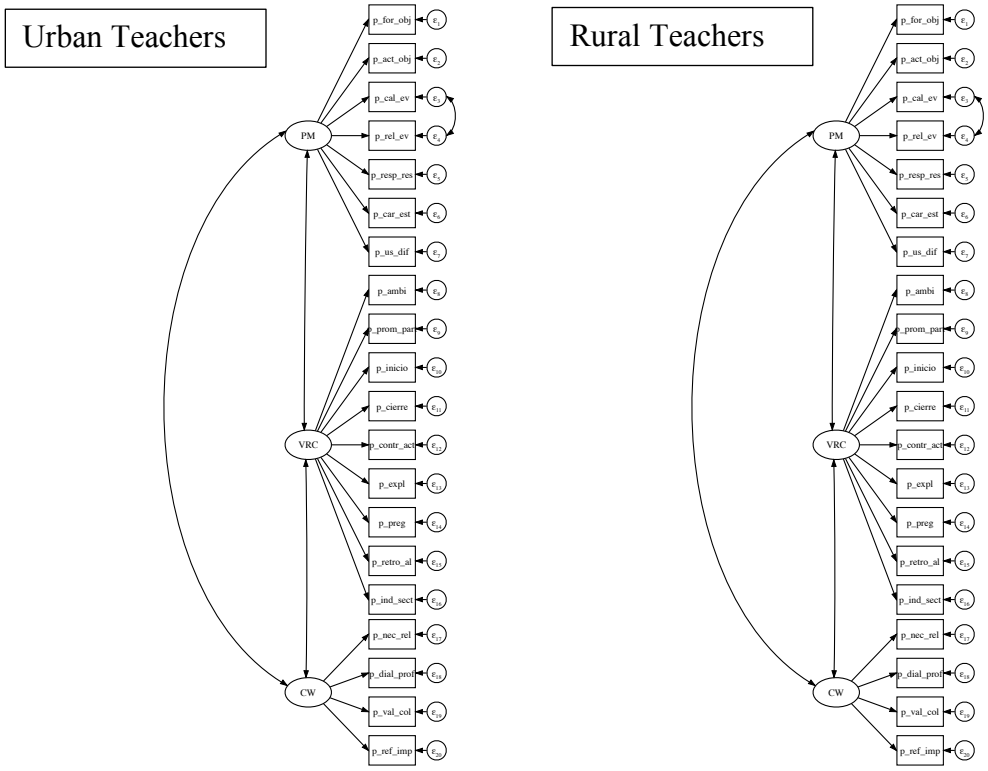


*Figure 4.10.* Hypothesizes multigroup baseline model of teacher evaluation portfolio for urban and rural teachers.

### The configural model

The configural model assumes that the factor structure is common among groups, i.e., in both groups, rural and urban teachers, the same items measure the same latent construct. Thus, the same items loaded onto the same factor across groups.

The results of the configural invariance model (Model 1) are presented in Table 4.14. The goal for this first model was to determine if the unconstrained multiple group meets the fit criteria. Therefore, the results for Model 1 were evaluated taking into consideration the cutoff criteria for three fit indexes: CFI, RMSEA, and SRMR (≥.95, ≤.06, and ≤.08, respectively; Hu & Bentler, 1995).

The fit indexes results for the configural model, suggested a well-fitting model ($\chi^2_{(332)} = 3{,}810.01$, $p < .001$; RMSEA = .031, CFI = .965, and SRMR = .029). These results for each one of the fit indexes were above or below the corresponding cutoff criteria. Thus, the results supported the presence of a three-factor model across urban and rural teachers, and the configural model became the model with which subsequent models were compared.

### The measurement model (weak factorial invariance model).

Since the configural invariance model was supported, the next step was to test for weak invariance (Model 2). In this model, I forced equal factor loading across the two groups, in addition to configural invariance. A common factor structure and loading are met if this model does not result in a deterioration of fit compared to the configural model (Model 1).

The results for Model 2 suggested a well-fitting model ($\chi^2_{(349)} = 3{,}635.04$, $p < .001$; RMSEA = .030, CFI = .967, and SRMR = .030), when they were compared with the cutoff criteria for three fix indexes. As can be seen in the table, the diff of chi-square was statistically significant ($p < .001$), indicating that Model 2 was significantly worse than the configural model. However, as has been previously mentioned, chi-square could be highly influenced by the sample size (Dimitrov, 2010). Therefore, evidence for ΔCFI, ΔRMSEA, and ΔSRMR were used in order to compare both models. The cutoff point of the change used for testing loading invariance was the proposed by Chen (2007): change equal to or lower than -.010 in CFI, supplemented by a change equal to or higher than .015 in RMSEA, or a change equal to or higher than .030 in SRMR indicating lack of invariance.

The results for Model 2, presented in Table 4.14, indicated that ΔCFI was .002, which is higher than the cutoff point of -.01 or lower. The ΔRMSEA was -.001, lower than the cutoff of .015, and for ΔSRMR the result was also -.001, lower than the cutoff criteria. The results, considering the changes in goodness of fit indexes, indicated that it is possible to accept the hypothesis that factor loadings are equal across rural and urban teachers.

The structural model (strong factorial invariance models)

This model adds the constraint of the identical threshold level going from one response category to the next for each indicator. As can be seen in Table 4.14, the diff of chi-square was statistically significant ($p < .001$), indicating that Model 3 is significantly worse than the weak model. On the other hand, the results for ΔCFI was -.003, a result

that although was negative, it was not lower than the cutoff criteria (-.01). ΔRMSEA and ΔSRMR were both lower than the cutoff criteria for the structural model (≥.015 and ≥.01). Therefore, considering the changes in goodness of fit indexes, the results indicated that as a function of the additional constraints of item threshold in the strong model, there were substantial improvements in model fit. According to these criteria, threshold invariances between both groups should be accepted.

Table 4.14.
*Goodness-of-Fit Statistics for Test of Measurement Invariance of a Three-Factor Model of the Teacher Evaluation Portfolio*

| Model | $\chi^2$ | (*df*) | CFI | RMSEA | SRMR | $\Delta\chi^2$ | (*df*) | Δ CFI | Δ RMSEA | Δ SRMR |
|---|---|---|---|---|---|---|---|---|---|---|
| Configural | 3,810.01 | (332) | .965 | .031 | .029 | | | | | |
| Measurement Model | 3,635.04 | (349) | .967 | .030 | .030 | 57.29*** | (17) | .002 | -.001 | -.001 |
| Structural Model | 3,966.33 | (406) | .964 | .029 | .030 | 343.3*** | (57) | -.003 | -.001 | .0 |

*Note*. RMSEA= Root mean square error of approximation; SRMR= Standardized root mean square residual; CIF= Comparative Fit Index;*** $p$ <.001

In summary, the results of measurement invariance considering urban and rural location indicated that the portfolio factor structures were invariant across these two subgroups of teachers. Therefore, this implies that for the Chilean Evaluation System, the indicators that comprise the portfolio operate equivalently across both populations of teachers, and the observed differences in the proportion of teachers evaluated as competent and outstanding can be attributed to differences in the teacher quality construct evaluated.

Descriptive statistics for teaching level

Table 4.15 describes the frequencies and percentages of the teacher performance level rating for the seven portfolio indicators of Module 1 depending on their teaching level: Early Childhood ($N$ = 1,459), Elementary School ($N$ = 4,250), Middle School ($N$ = 6,875), High School ($N$ = 4,137), Special Education ($N$ = 4,265), and Adult Education ($N$ = 442; missing = 550). The distribution of teachers depending on their performance level was the same that was previously described for the whole sample of teachers.

When I compared the differences in the distribution of the teachers depending on their teaching level, significant differences in the proportion of teachers related to their performance level were found for all the indicators of Module 1 (Ind.1.1 $\chi^2_{(15)}$=283.75, $p < 0.01$; Ind.1.2: $\chi^2_{(15)}$=732.44, $p < 0.01$; Ind.2.1: $\chi^2_{(15)}$=278.33, $p < 0.01$; Ind.2.2: $\chi^2_{(15)}$=346.93, $p < 0.01$; Ind.2.3: $\chi^2_{(15)}$=342.39, $p < 0.01$; Ind.3.1: $\chi^2_{(15)}$=262.82, $p < 0.01$; Ind.3.2: $\chi^2_{(15)}$=154.49, $p < 0.01$).

Table 4.15.

*Frequencies and percentages of the 7 Module 1 portfolio indicators for teachers based on the level taught*

| Indicator | Early Childhood | | Elementary | | Middle School | | High School | | Special Education | | Adult Education | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Frequency | Percent | Frequency | Percent | Frequency | Percent | Frequency | Percent | Frequency | Percent | Frequency | Percent |
| 1.1.Formation of learning objectives *** | | | | | | | | | | | | |
| Unsatisfactory | 22 | 1.51 | 48 | 1.13 | 188 | 2.73 | 125 | 3.02 | 65 | 1.52 | 20 | 4.52 |
| Basic | 73 | 5.00 | 237 | 5.58 | 624 | 9.08 | 377 | 9.11 | 290 | 6.80 | 35 | 7.92 |
| Competent | 1,234 | 84.58 | 3,212 | 75.58 | 5,084 | 73.95 | 3,188 | 77.06 | 3,309 | 77.58 | 363 | 82.13 |
| Outstanding | 130 | 8.91 | 753 | 17.72 | 979 | 14.24 | 447 | 10.80 | 601 | 14.09 | 24 | 5.43 |
| 1.2.Relationship between activities and objectives *** | | | | | | | | | | | | |
| Unsatisfactory | 190 | 13.02 | 374 | 8.80 | 1,222 | 17.77 | 791 | 19.12 | 353 | 8.28 | 97 | 21.95 |
| Basic | 372 | 25.50 | 753 | 17.72 | 1,646 | 23.94 | 1,080 | 26.11 | 795 | 18.64 | 134 | 30.32 |
| Competent | 856 | 58.67 | 2,969 | 69.86 | 3,880 | 56.44 | 2,219 | 53.64 | 3,012 | 70.62 | 208 | 47.06 |
| Outstanding | 41 | 2.81 | 154 | 3.62 | 127 | 1.85 | 47 | 1.14 | 105 | 2.46 | 3 | 0.68 |
| 2.1. Evaluation and rubrics used for correction *** | | | | | | | | | | | | |
| Unsatisfactory | 267 | 18.50 | 502 | 11.92 | 1,078 | 15.89 | 847 | 20.66 | 752 | 17.77 | 64 | 14.71 |
| Basic | 485 | 33.61 | 2,225 | 52.83 | 3,030 | 44.67 | 1,781 | 43.45 | 1,883 | 44.48 | 207 | 47.59 |
| Competent | 681 | 47.19 | 1,428 | 33.90 | 2,609 | 38.46 | 1,427 | 34.81 | 1,536 | 36.29 | 159 | 36.55 |
| Outstanding | 10 | 0.69 | 57 | 1.35 | 66 | 0.97 | 44 | 1.07 | 62 | 1.46 | 5 | 1.15 |
| 2.2.Relationship between assessment and objectives *** | | | | | | | | | | | | |
| Unsatisfactory | 301 | 20.86 | 512 | 12.15 | 1,224 | 18.04 | 813 | 19.84 | 661 | 15.62 | 80 | 18.39 |
| Basic | 503 | 34.86 | 1,506 | 35.74 | 2,438 | 35.93 | 1,314 | 32.07 | 1,385 | 32.72 | 135 | 31.03 |
| Competent | 592 | 41.03 | 1,806 | 42.86 | 2,792 | 41.15 | 1,828 | 44.62 | 1,804 | 42.62 | 208 | 47.82 |
| Outstanding | 47 | 3.26 | 390 | 9.25 | 331 | 4.88 | 142 | 3.47 | 383 | 9.05 | 12 | 2.76 |

*p<0.05; **p<0.01;***p<0.001 (Chi-square test)

Note: Chi-square test with significance levels used for differences between the teachers for six different teaching levels.

| Indicator | Early Childhood | | Elementary | | Middle School | | High School | | Special Education | | Adult Education | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Frequency | Percent | Frequency | Percent | Frequency | Percent | Frequency | Percent | Frequency | Percent | Frequency | Percent |
| 2.3.Analysis and use of assessment results *** | | | | | | | | | | | | |
| Unsatisfactory | 312 | 21.64 | 548 | 13.04 | 1,401 | 20.71 | 928 | 22.75 | 636 | 15.06 | 117 | 26.96 |
| Basic | 469 | 32.52 | 1,576 | 37.49 | 2,336 | 34.53 | 1,359 | 33.32 | 1,569 | 37.16 | 159 | 36.64 |
| Competent | 637 | 44.17 | 1,967 | 46.79 | 2,884 | 42.62 | 1,729 | 42.39 | 1,801 | 42.66 | 150 | 34.56 |
| Outstanding | 24 | 1.66 | 113 | 2.69 | 145 | 2.14 | 63 | 1.54 | 216 | 5.12 | 8 | 1.84 |
| 3.1.Analyses based on students' characteristics *** | | | | | | | | | | | | |
| Unsatisfactory | 171 | 11.75 | 425 | 10.02 | 661 | 9.67 | 449 | 10.90 | 439 | 10.31 | 54 | 12.30 |
| Basic | 818 | 56.22 | 1,903 | 44.88 | 2,949 | 43.12 | 1,546 | 37.54 | 2,069 | 48.60 | 169 | 38.50 |
| Competent | 448 | 30.79 | 1,792 | 42.26 | 3,077 | 44.99 | 2,049 | 49.76 | 1,642 | 38.57 | 209 | 47.61 |
| Outstanding | 18 | 1.24 | 120 | 2.83 | 152 | 2.22 | 74 | 1.80 | 107 | 2.51 | 7 | 1.59 |
| 3.2.Use of error for learning *** | | | | | | | | | | | | |
| Unsatisfactory | 184 | 12.66 | 319 | 7.54 | 812 | 11.90 | 407 | 9.91 | 298 | 7.02 | 47 | 10.76 |
| Basic | 348 | 23.95 | 1,015 | 24.00 | 1,744 | 25.57 | 1,068 | 26.00 | 1,091 | 25.70 | 115 | 26.32 |
| Competent | 881 | 60.63 | 2,759 | 65.22 | 4,122 | 60.43 | 2,549 | 62.05 | 2,701 | 63.63 | 268 | 61.33 |
| Outstanding | 40 | 2.75 | 137 | 3.24 | 143 | 2.10 | 84 | 2.04 | 155 | 3.65 | 7 | 1.60 |

*p<0.05; **p<0.01;***p<0.001 (Chi-square test)

Note: Chi-square test with significance levels used for differences between the teachers for six different teaching levels.

Since the omnibus chi-square value does not specify which combination of categories contributes to statistical significance, a post-hoc comparison analysis for interpreting the differences of the chi-square contingency tables was carried out using the standardized residual method (Beasley & Schumacker, 1995). In order to follow the method, I calculated the chi-square associated with each cell, and then I estimated *p* values that were compared against the Bonferroni corrected *p*-value ($\alpha$=0.05/ 24[14]= 0.0021).

Using this methodology, I observed for the Module 1 indicators, that in general teachers who taught early childhood, middle school, high school, and adult education showed a significantly lower proportion of teachers as competent or outstanding than the teachers who taught in elementary or special education levels.

[14] 24 corresponds to the possible group comparisons (6 different groups with 4 possible levels).

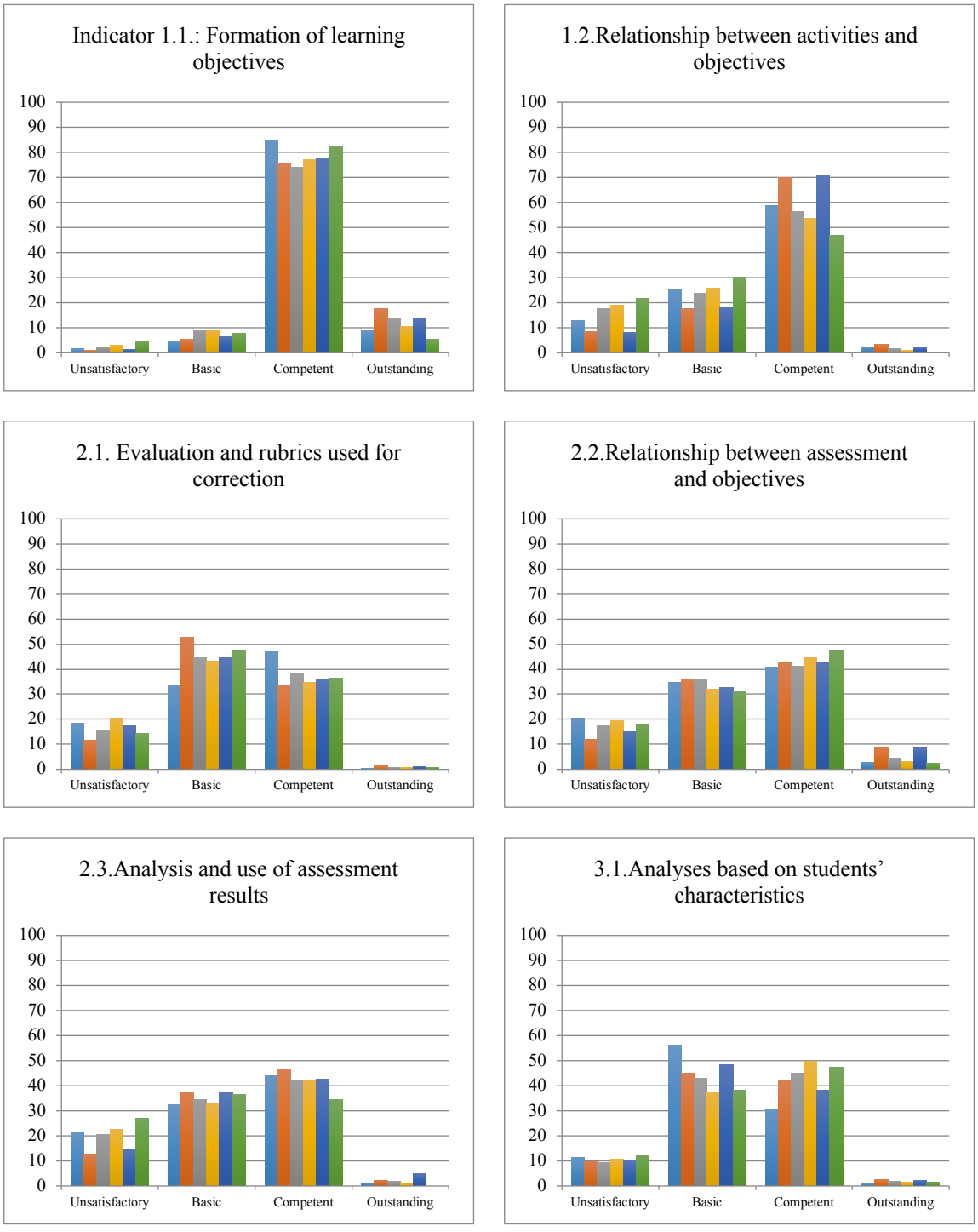


*Figure 4.11.* Percent of responses for the 7 Module 1 portfolio indicators for teachers based on the level taught.

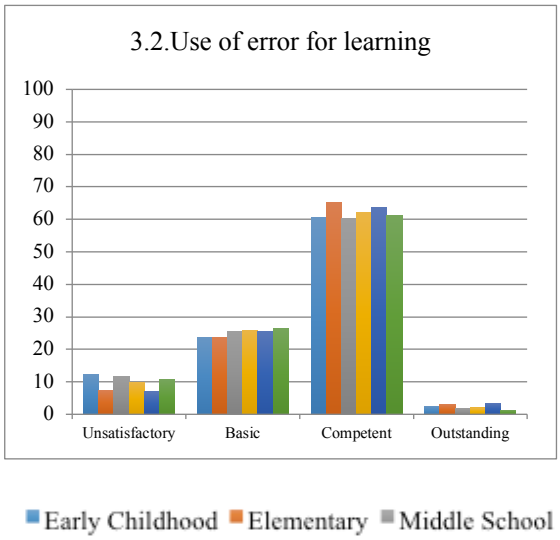


Table 4.16 describes the frequencies and percentages of the teacher performance level ratings for the nine portfolio indicators of Module 2, depending on the teaching level. As can be seen in the table, the distribution of teachers at each performance level for Module 2 was the same that was described for the whole sample of teachers for aim 1.

Considering the different teaching levels, significant differences in the proportion of teachers in each category were found for all the indicators of Module 2 (Ind.4.1 $\chi^2_{(15)}$=419.16, $p < 0.01$; Ind.4.2: $\chi^2_{(15)}$=345.43, $p < 0.01$; Ind.4.3: $\chi^2_{(15)}$=717.16, $p < 0.01$; Ind.4.4: $\chi^2_{(15)}$=681.17, $p < 0.01$; Ind.4.5: $\chi^2_{(15)}$=5,058.35, $p < 0.01$; Ind. 4.5: $\chi^2_{(15)}$= 1,334.96, $p < 0.01$; Ind.4.7: $\chi^2_{(15)}$=1,100.23, $p < 0.01$; Ind.4.8: $\chi^2_{(15)}$= 959.96, $p < 0.01$; Ind.4.9: $\chi^2_{(15)}$= 836.54, $p < 0.01$). Overall, post-hoc comparison analysis indicates that teachers who teach early childhood and elementary school showed a significantly lower proportion of competent or outstanding teachers than high school or special education teachers. Middle school or adult education teachers did not show as much of a significant difference.

Table 4.16.

*Frequencies and percentages of the 9 Module 2 portfolio indicators for teachers based on the level taught*

| Indicator | Early Childhood | | Elementary | | Middle School | | High School | | Special Education | | Adult Education | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Frequency | Percent | Frequency | Percent | Frequency | Percent | Frequency | Percent | Frequency | Percent | Frequency | Percent |
| 4.1.Class environment *** | | | | | | | | | | | | |
| Unsatisfactory | 17 | 1.17 | 47 | 1.11 | 82 | 1.19 | 40 | 0.97 | 24 | 0.56 | 2 | 0.45 |
| Basic | 65 | 4.46 | 331 | 7.79 | 416 | 6.05 | 143 | 3.46 | 110 | 2.58 | 13 | 2.93 |
| Competent | 1,339 | 91.84 | 3,834 | 90.21 | 6,038 | 87.86 | 3,737 | 90.35 | 3,792 | 88.93 | 407 | 91.87 |
| Outstanding | 37 | 2.54 | 38 | 0.89 | 336 | 4.89 | 216 | 5.22 | 338 | 7.93 | 21 | 4.74 |
| 4.2.Quality of the start of class *** | | | | | | | | | | | | |
| Unsatisfactory | 127 | 8.71 | 130 | 3.06 | 335 | 5.17 | 168 | 4.06 | 175 | 4.10 | 8 | 1.81 |
| Basic | 521 | 35.73 | 1,361 | 32.02 | 1,613 | 23.47 | 1,031 | 24.93 | 1,032 | 24.20 | 117 | 26.41 |
| Competent | 800 | 54.87 | 2,740 | 64.47 | 4,767 | 69.37 | 2,861 | 69.17 | 3,015 | 70.71 | 316 | 71.33 |
| Outstanding | 10 | 0.69 | 19 | 0.45 | 137 | 1.99 | 76 | 1.84 | 42 | 0.98 | 2 | 0.45 |
| 4.3.Quality of the end of class *** | | | | | | | | | | | | |
| Unsatisfactory | 256 | 17.56 | 195 | 4.59 | 650 | 9.46 | 324 | 7.83 | 231 | 5.42 | 24 | 5.42 |
| Basic | 815 | 55.90 | 1,998 | 47.01 | 2,968 | 43.19 | 1,675 | 40.50 | 1,914 | 44.89 | 203 | 45.82 |
| Competent | 373 | 25.58 | 2,034 | 47.86 | 3,050 | 44.38 | 1,943 | 46.98 | 1,882 | 44.14 | 208 | 46.95 |
| Outstanding | 140 | 0.96 | 23 | 0.54 | 204 | 2.97 | 194 | 4.69 | 237 | 5.56 | 8 | 1.81 |
| 4.4.Contribution of the activities to the achievement of the class objectives *** | | | | | | | | | | | | |
| Unsatisfactory | 124 | 8.52 | 212 | 4.99 | 939 | 13.67 | 503 | 12.16 | 167 | 3.92 | 55 | 12.42 |
| Basic | 136 | 9.34 | 877 | 20.64 | 1,339 | 19.49 | 663 | 16.03 | 611 | 14.33 | 57 | 12.87 |
| Competent | 1,177 | 80.84 | 3,130 | 73.65 | 4,447 | 64.72 | 2,808 | 69.65 | 3,421 | 80.23 | 326 | 73.59 |
| Outstanding | 19 | 1.30 | 31 | 0.73 | 146 | 2.12 | 89 | 2.15 | 65 | 1.52 | 5 | 1.13 |

*p<0.05; **p<0.01;***p<0.001 (Chi-square test)

Note: Chi-square test with significance levels used for differences between the teachers for six different teaching levels.

| Indicator | Early Childhood | | Elementary | | Middle School | | High School | | Special Education | | Adult Education | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Frequency | Percent | Frequency | Percent | Frequency | Percent | Frequency | Percent | Frequency | Percent | Frequency | Percent |
| 4.5.Curricular emphasis on the subject *** | | | | | | | | | | | | |
| Unsatisfactory | 490 | 33.61 | 1,587 | 37.34 | 859 | 12.50 | 493 | 11.92 | 2,720 | 62.79 | 100 | 22.57 |
| Basic | 542 | 37.17 | 2,243 | 52.78 | 3,774 | 54.92 | 2,037 | 49.25 | 1,322 | 31.00 | 211 | 47.63 |
| Competent | 423 | 29.01 | 419 | 9.86 | 2,184 | 31.78 | 1,495 | 36.15 | 206 | 4.83 | 128 | 28.89 |
| Outstanding | 3 | 0.21 | 1 | 0.02 | 55 | 0.80 | 111 | 2.68 | 16 | 0.38 | 4 | 0.90 |
| 4.6.Clear explanations *** | | | | | | | | | | | | |
| Unsatisfactory | 389 | 26.68 | 239 | 5.62 | 723 | 10.52 | 281 | 6.79 | 202 | 4.74 | 16 | 3.61 |
| Basic | 820 | 56.24 | 2,922 | 68.75 | 3,704 | 53.90 | 2,091 | 50.56 | 2,201 | 51.62 | 259 | 58.47 |
| Competent | 247 | 16.94 | 1,085 | 25.53 | 2,406 | 35.01 | 1,698 | 41.05 | 1,814 | 42.54 | 165 | 37.25 |
| Outstanding | 2 | 0.14 | 4 | 0.09 | 39 | 0.57 | 66 | 1.60 | 47 | 1.10 | 3 | 0.68 |
| 4.7.Questions and activities *** | | | | | | | | | | | | |
| Unsatisfactory | 21 | 1.44 | 30 | 0.71 | 69 | 1.00 | 45 | 1.09 | 362 | 0.84 | 9 | 2.03 |
| Basic | 1,150 | 78.88 | 3,033 | 71.36 | 4,332 | 63.04 | 2,353 | 56.89 | 1,910 | 44.79 | 258 | 58.24 |
| Competent | 280 | 19.20 | 1,170 | 27.53 | 2,300 | 33.47 | 1,538 | 37.19 | 2,020 | 47.37 | 165 | 37.25 |
| Outstanding | 7 | 0.48 | 17 | 0.40 | 171 | 2.49 | 200 | 4.84 | 298 | 6.99 | 11 | 2.48 |
| 4.8.Encouragement *** | | | | | | | | | | | | |
| Unsatisfactory | 15 | 1.03 | 29 | 0.68 | 580 | 0.84 | 35 | 0.85 | 17 | 0.40 | 5 | 1.13 |
| Basic | 792 | 54.32 | 1,470 | 34.59 | 1,668 | 24.27 | 922 | 22.29 | 1,592 | 37.34 | 138 | 31.15 |
| Competent | 628 | 43.07 | 2,688 | 63.25 | 4,621 | 67.24 | 2,904 | 70.21 | 2,415 | 56.64 | 284 | 64.11 |
| Outstanding | 23 | 1.58 | 63 | 1.48 | 525 | 7.64 | 275 | 6.65 | 240 | 5.63 | 16 | 3.61 |
| 4.9.Student feedback *** | | | | | | | | | | | | |
| Unsatisfactory | 277 | 19.00 | 533 | 12.54 | 481 | 7.00 | 291 | 7.04 | 357 | 8.37 | 34 | 7.67 |
| Basic | 702 | 48.49 | 2,011 | 47.32 | 4,295 | 62.50 | 2,583 | 62.45 | 2,179 | 51.10 | 308 | 69.53 |
| Competent | 450 | 30.86 | 1,615 | 38.00 | 1,905 | 27.72 | 1,152 | 27.85 | 1,407 | 33.00 | 97 | 21.90 |
| Outstanding | 24 | 1.65 | 91 | 2.14 | 191 | 2.78 | 110 | 2.66 | 321 | 7.53 | 4 | 0.90 |

*p<0.05; **p<0.01;***p<0.001 (Chi-square test)

Note: Chi-square test with significance levels used for differences between the teachers for six different teaching levels.


*Figure 4.12.* Percentages of responses for the 9 Module 2 portfolio indicators for teachers based on the level taught

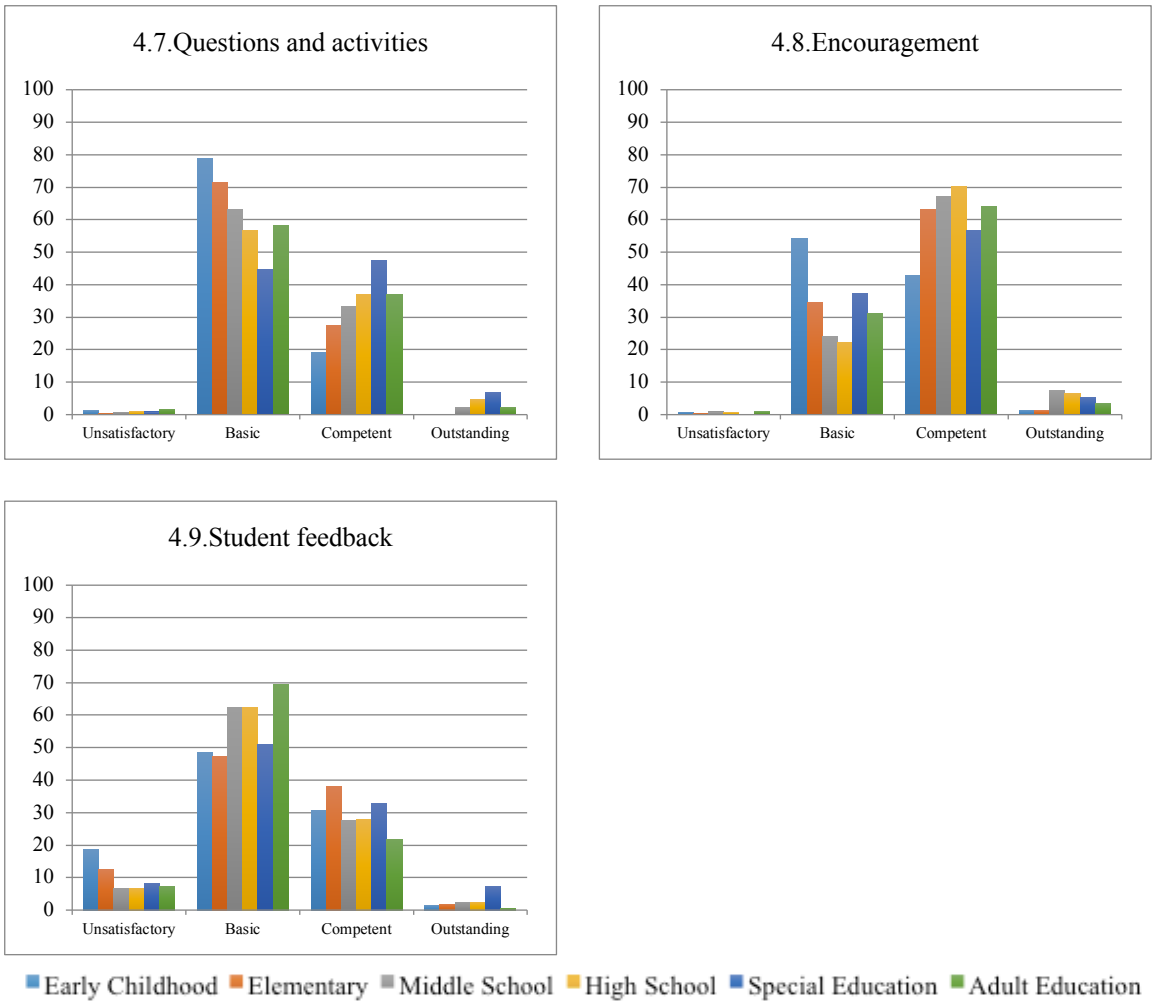


*Figure 4.13.* Percentages of responses for the 9 Module 2 portfolio indicators for teachers based on the level taught

Table 4.17 describes the frequencies and percentages of the teacher performance level for different teaching levels within the four portfolio indicators of Module 3. As can be seen in the table, the distribution of teachers depending on their performance level for Module 3 was the same as previously described for the whole sample of teachers in aim 1.

Based on the teaching level, significant differences in the proportion of teachers' performance level were found for all of the indicators in Module 3 (Ind.5.1 $\chi^2_{(15)} = 179.17$, $p < 0.01$; Ind.5.2: $\chi^2_{(15)} = 336.1$, $p < 0.01$; Ind.5.3: $\chi^2_{(15)} = 202.6$, $p < 0.01$; Ind.5.4: $\chi^2_{(15)} = 378.69$, $p < 0.01$). Post-hoc comparison analysis indicated that within the Module 3 indicators, teachers who teach early childhood and middle school showed a significantly lower proportion of competent or outstanding compared to elementary or special education teachers. High school and adult education teachers did not show such a significant difference as the other categories.

Table 4.17.

*Frequencies and percentages of the 4 Module 3 portfolio indicators for teachers based on the level taught*

| Indicator | Early Childhood | | Elementary | | Middle School | | High School | | Special Education | | Adult Education | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Frequency | Percent | Frequency | Percent | Frequency | Percent | Frequency | Percent | Frequency | Percent | Frequency | Percent |
| 5.1.Collaborative work suitability *** | | | | | | | | | | | | |
| Unsatisfactory | 151 | 11.60 | 363 | 9.74 | 699 | 12.44 | 381 | 11.95 | 295 | 7.65 | 41 | 15.47 |
| Basic | 618 | 47.47 | 1,564 | 41.98 | 2,345 | 41.75 | 1,207 | 37.87 | 1,697 | 44.01 | 108 | 40.75 |
| Competent | 452 | 34.72 | 1,343 | 36.04 | 2,147 | 38.22 | 1,284 | 40.29 | 1,434 | 37.19 | 96 | 36.23 |
| Outstanding | 81 | 6.22 | 456 | 12.24 | 426 | 7.58 | 315 | 9.88 | 430 | 11.15 | 20 | 7.55 |
| 5.2.Quality of professional dialogue*** | | | | | | | | | | | | |
| Unsatisfactory | 267 | 20.59 | 590 | 15.89 | 1,175 | 20.97 | 597 | 18.84 | 502 | 13.05 | 62 | 23.40 |
| Basic | 539 | 41.56 | 1,008 | 27.14 | 1,687 | 30.11 | 1,018 | 32.13 | 1,464 | 38.05 | 91 | 34.34 |
| Competent | 427 | 32.92 | 1,665 | 44.83 | 2,303 | 41.11 | 1,290 | 40.72 | 1,514 | 39.35 | 103 | 38.87 |
| Outstanding | 64 | 4.93 | 451 | 12.14 | 437 | 7.80 | 263 | 8.30 | 368 | 9.56 | 9 | 3.40 |
| 5.3.Value of collaborative work for professional development *** | | | | | | | | | | | | |
| Unsatisfactory | 497 | 38.50 | 1,139 | 30.81 | 2,158 | 38.79 | 1,158 | 36.75 | 1,399 | 36.50 | 107 | 40.84 |
| Basic | 502 | 38.88 | 1,375 | 37.19 | 1,698 | 30.52 | 875 | 27.77 | 1,097 | 28.62 | 74 | 28.24 |
| Competent | 284 | 22.00 | 1,117 | 30.21 | 1,654 | 29.73 | 1,071 | 33.99 | 1,273 | 33.21 | 78 | 29.77 |
| Outstanding | 8 | 0.62 | 66 | 1.79 | 54 | 0.97 | 47 | 1.49 | 64 | 1.67 | 3 | 1.15 |
| 5.4. Reflection on the impact of the collaborative work experience *** | | | | | | | | | | | | |
| Unsatisfactory | 289 | 22.39 | 501 | 13.55 | 1,039 | 18.67 | 550 | 17.45 | 563 | 14.69 | 54 | 20.61 |
| Basic | 723 | 56.00 | 1,670 | 45.17 | 2,429 | 43.66 | 1,396 | 44.30 | 1,962 | 51.19 | 127 | 48.47 |
| Competent | 275 | 21.30 | 1,355 | 36.65 | 1,992 | 35.80 | 1,135 | 36.02 | 1,117 | 29.14 | 77 | 29.39 |
| Outstanding | 4 | 0.31 | 171 | 4.63 | 104 | 1.87 | 70 | 2.22 | 191 | 4.98 | 4 | 1.53 |

*p<0.05; **p<0.01;***p<0.001 (Chi-square test)

Note: Chi-square test with significance levels used for differences between the teachers for six different teaching levels.

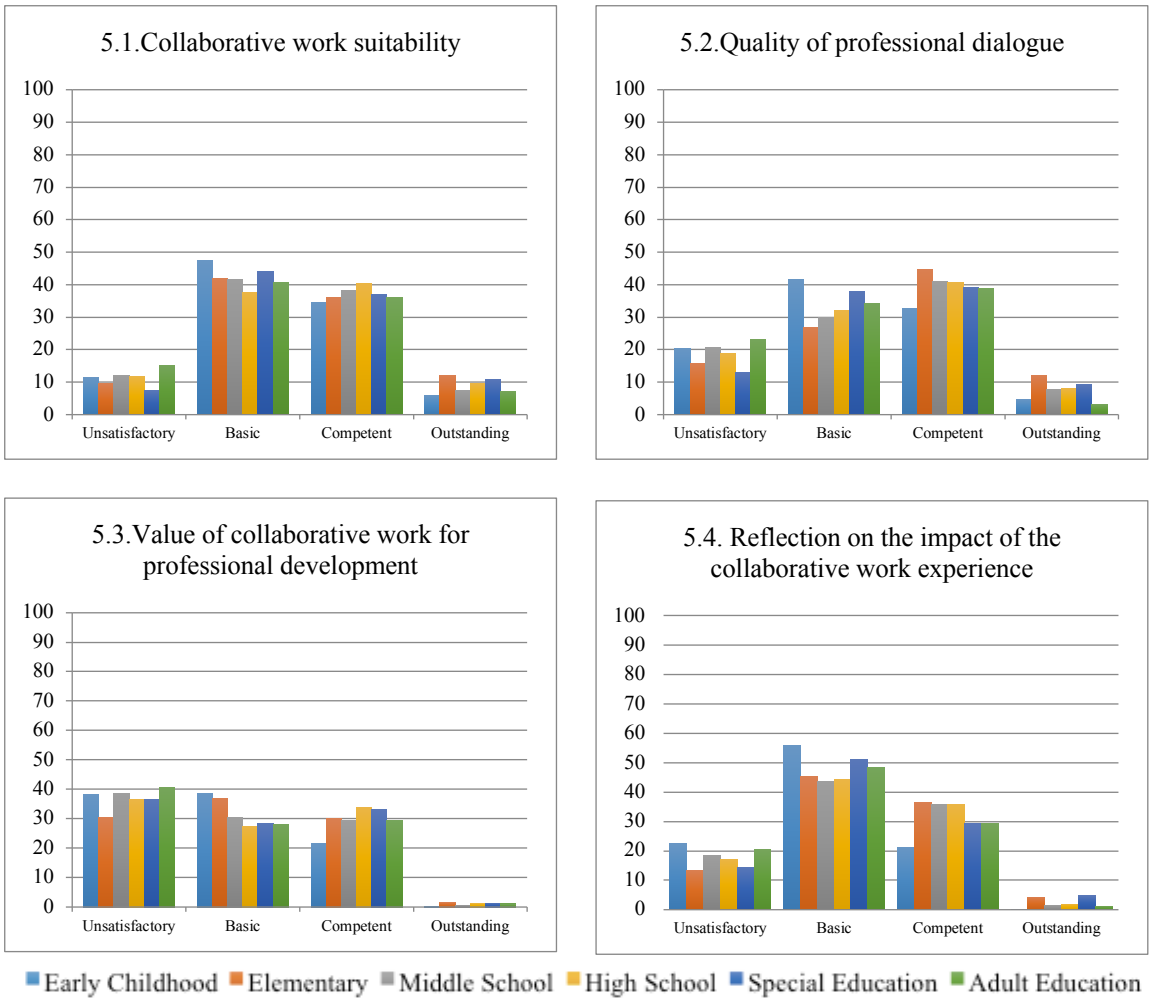


*Figure 4.14.* Percentage of responses of the 4 Module 3 portfolio indicators for teachers based on the level taught.

## Measurement invariance for teaching level

Measurement invariance for teachers based on the teaching level was also tested within the framework of multigroup confirmatory factor analysis (CFA). Each one of the steps of measurement invariance are presented below.

### The Hypothesized Model.

A well-fitting baseline model structure was established for each one of the six groups of teachers based on their teaching level. Once these models were established, they represented the hypothesized multigroup model under test. The model under test in this initial multigroup measurement of variance was the same one that proposed a three-factor portfolio structure solution, with a covariance between indicators 2.1 and 2.2, which was previously tested in aim 1 for all of the teachers evaluated. This structure served as the initial model tested in the establishment of baseline models for the six groups of teachers.

***Establishing Baseline Models***. Based on the teaching level, there were six groups analyzed in the current research. For each group, six separate CFA were conducted evaluating a three-factor structure of the portfolio with one residual covariance between indicators 2.1 and 2.2. Adequate goodness-of-fit results were reported for each separate analysis, with the exception of the baseline model for high school teachers (group 4):

Early childhood teachers: $\chi^2_{(166)} = 475.6$, $p < .001$; RMSEA = .036, CFI = .951.

Elementary school teachers: $\chi^2_{(166)} = 836.49$, $p < .001$; RMSEA = .031, CFI = .965.

Middle school teachers: $\chi^2_{(166)} = 836.49$, $p < .001$; RMSEA = .031, CFI = .965.

High school teachers: $\chi^2_{(166)} = 1{,}183.07$, $p < .001$; RMSEA = .038, **CFI = .945**.

Special education teachers: $\chi^2_{(166)} = 900.32$, $p < .001$; RMSEA = .032, CFI = .967.

Adult education teachers: $\chi^2_{(166)} = 235.54$, $p < .001$; RMSEA = .031, CFI = .959.

In order to improve the high school teacher baseline model, the MI that was substantially larger than all of the other MIs was reviewed. The residual covariance between indicators 4.2 and 4.3 was selected to incorporate it into a new post-hoc model for high school teachers. Results from the estimation of this new model for high school teachers yielded satisfactory goodness-of-fit statistics: $\chi^2_{(165)} = 1{,}067.92$, $p < .001$; RMSEA = .036, CFI = .951. Therefore, this modified model served as the baseline for high school teachers.

Figure 4.11 shows the representation of the baseline models for the six groups of teachers depending on their level taught, and it provides the foundation with which I tested the series of stringent between-group constraints related to portfolio structure.

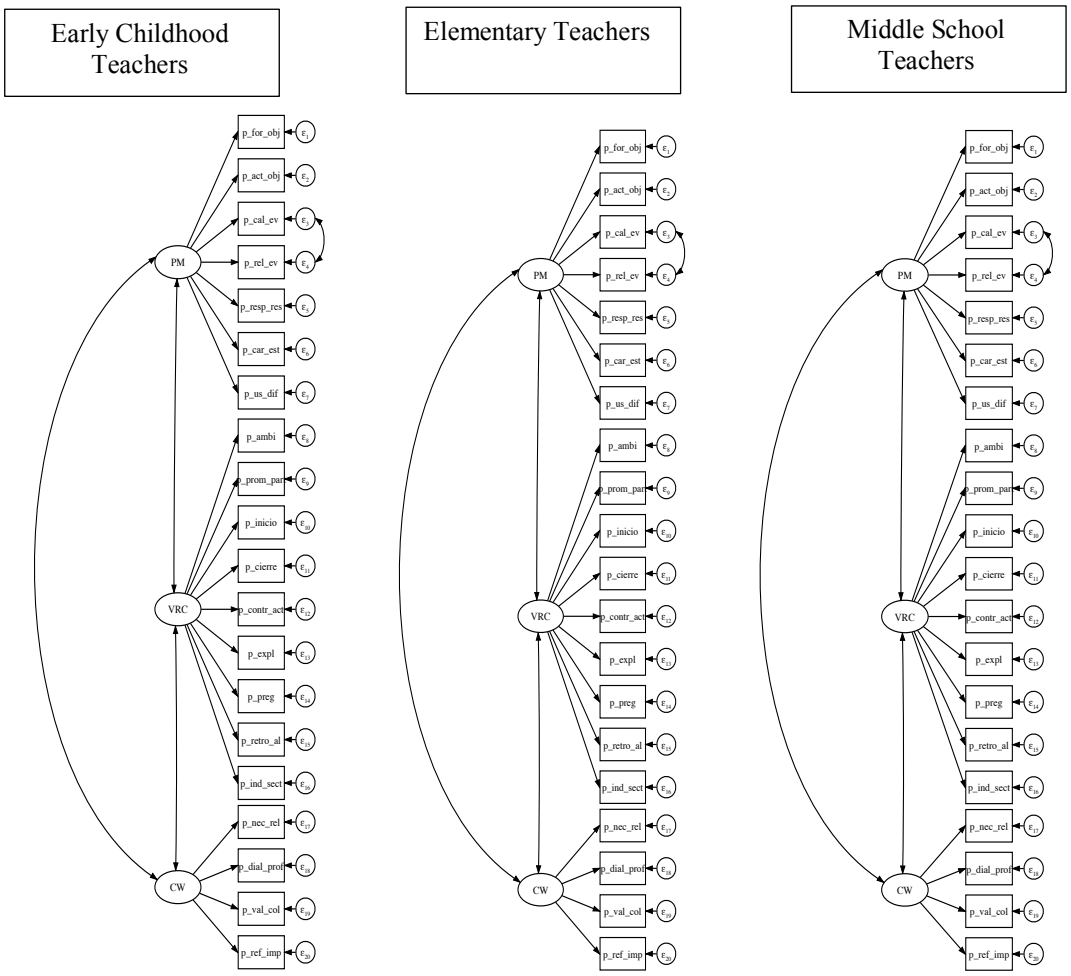


*Figure 4.15*. Hypothesizes multigroup baseline models of teacher evaluation portfolio based on teaching level.

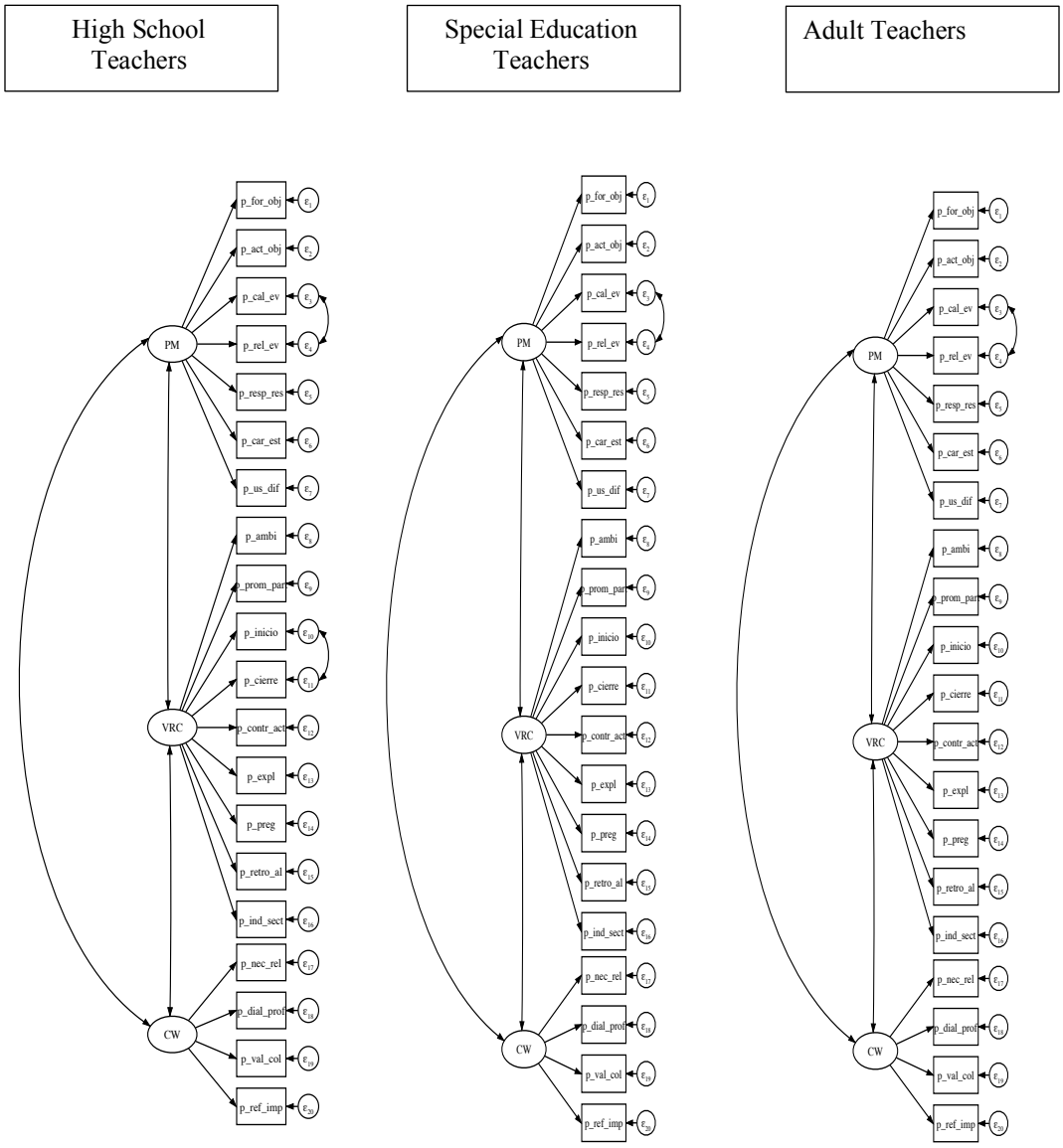


*Figure 4.16.* Hypothesizes multigroup baseline models of teacher evaluation portfolio based on teaching level.

### The configural model

Results for the configural invariance model of teaching level (Model 1) are presented in Table 4.18. Taking into consideration the cutoff criteria for the fit indexes CFI, RMSEA, and SRMR, the results for the configural model indicated a well-fitting model ($\chi^2_{(995)} = 5{,}254.62$, $p < .001$; RMSEA = .035, CFI = .957, and SRMR = .035). Therefore, the results supported a common three-factor structure across the six groups of teaching levels, and the configural model became the model with which subsequent models were compared.

### The measurement model (weak factorial invariance models).

Results for the weak model suggested a well-fitting model ($\chi^2_{(1,080)} = 5{,}538.35$, $p < .001$; RMSEA = .034, CFI = .955, and SRMR = .039). As can be seen in Table 4.18, the diff of chi-square was statistically significant ($p < .001$), indicating that Model 2 is significantly worse than the configural model. However, as was indicated before, chi-square could be highly influenced by the sample size, therefore evidence for ΔCFI, ΔRMSEA, and ΔSRMR were used in order to compare both models. In order to compare both models the cutoff point of the change used was the one proposed by Chen (2007) for the weak factorial invariance model: change of equal to or lower than -.010 in CFI, supplemented by a change equal to or higher than .015 in RMSEA, or a change equal to or higher than .030 in SRMR indicating lack of invariance.

The results for the weak model, presented in Table 4.18, indicated that ΔCFI was -.002, which is higher than the cutoff point of -.01. The ΔRMSEA was -.001, lower than the cutoff of .015, and for ΔSRMR the result was .004, also lower than the cutoff criteria.

The results, considering the changes in goodness of fit indexes, indicated that it is possible to accept the hypothesis that factor loadings are equal across the six groups of teachers.

The structural model (strong factorial invariance models)

This model adds the constraint of the identical threshold level going from one response category to the next for each indicator. As can be seen in Table 4.18, the diff of chi-square was statistically significant ($p$ <.001), indicating that Model 3 is significantly worse than the weak model. Additionally, the results for ΔCFI was -.12, resulting lower than the cutoff criteria (-.01). ΔRMSEA was .023, higher than the cutoff criteria for the structural model (≥.015), and ΔSRMR was .007, slightly lower than the cutoff (≥.01) Therefore, considering the changes in CFI and RMSEA, the results indicated that as a function of the additional constraints of item threshold in the strong model, there were no substantial improvements in model fit. According to these criteria, threshold invariances between the six levels should be rejected.

The results of the measurement invariance considering six teaching levels, indicated that the portfolio factor structures were invariant at the structure and factor loadings. However, for the threshold from one category of performance to the other, the portfolio was not equivalent in the six teacher groups.

Table 4.18.

*Goodness-of-Fit Statistics for Test of Measurement Invariance of a Three-Factor Model of the Teacher Evaluation Portfolio*

| Model | $\chi^2$ | (*df*) | CFI | RMSEA | SRMR | $\Delta\chi^2$ | (*df*) | Δ CFI | Δ RMSEA | Δ SRMR |
|---|---|---|---|---|---|---|---|---|---|---|
| Configural | 5,254.6 | (995) | .957 | .035 | .035 | | | | | |
| Measurement Model | 5,538.4 | (1,080) | .955 | .034 | .039 | 510.5*** | (85) | -.002 | -.001 | .004 |
| Structural Model | 17,346.2 | (1.366) | .840 | .057 | .046 | 13.731*** | (286) | -.12 | .023 | .007 |

*Note*. RMSEA= Root mean square error of approximation; SRMR= Standardized root mean square residual; CIF= Comparative Fit Index;

*** $p$ <.001

Aim 3: Compare the theoretical weight assigned to each one of the portfolio indicators to the empirical data from the Chilean Teacher Evaluation System.

As was described in Table 1.1, each indicator within the portfolio has a different weight regarding the importance of the indicator as an approach to teacher quality. Therefore, for the teachers whose Module 3 was included in their final score, each of the seven indicators that are part of Module 1 weighed 4%. The same weight was given to each one of the four indicators in Module 3. For Module 2, five indicators (*class environment*, *quality of the start of class*, *quality of the end of class*, *the contribution of the activities to the achievement of the class objectives*, and *encouragement*) weighed 4%, and the other four remaining indicators (*curricular emphasis on the subject, clear explanations, questions and activities, and student feedback and use of assessment results*) weighed 9%. Likewise, for the teachers whose Module 3 was not included in their final score, Module 1 indicators weighed 5%. For Module 2, the indicators that previously weighed 4% changed to a weight of 5%, and the 9% weight changed to10%.

In order to assess the empirical weights, I used a weighted sum score method. With this method, the sum score can be obtained when the factor loading of each item is multiplied to the scaled score for each item before summing. Thus, the first step was to inspect the factor loading for the ECFA 3-factor solution when M3 was included in the final score, and for the ECFA 2-factor solution when M3 was not included in the final score. Tables 4.19 and 4.20 show the factor loading results for each model. The results for the factor loading are presented in descending order in both tables, in order to inspect if the highest factorial loads actually correspond to the highest theoretical weights.

Table 4.19.
*Exploratory Confirmatory Factor Analysis loadings of the portfolio indicators from Teacher Evaluation 2017 with M3 results (N=11,216)*

| | Indicator | Factor loading | Factor | Weight |
|---|---|---|---|---|
| 1 | 2.2.Relationship between assessment and objectives | 0.915 | 1 | 4% |
| **2** | **4.7.Questions and activities** | **0.702** | **2** | **9%** |
| **3** | **4.6.Clear explanations** | **0.647** | **2** | **9%** |
| 4 | 4.3.Quality of the end of class | 0.630 | 2 | 4% |
| 5 | 2.1. Evaluation and rubrics used for correction | 0.624 | 1 | 4% |
| 6 | 4.2.Quality of the start of class | 0.591 | 2 | 4% |
| 7 | 3.1.Analyses based on students' characteristics | 0.581 | 1 | 4% |
| 8 | 3.2.Use of error for learning | 0.560 | 1 | 4% |
| **9** | **4.9.Student feedback** | **0.556** | **2** | **9%** |
| 10 | 5.4. Reflection on the impact of the collaborative work experience | 0.521 | 3 | 4% |
| 11 | 4.1.Class environment | 0.488 | 2 | 4% |
| 12 | 5.3.Value of collaborative work for professional development | 0.464 | 3 | 4% |
| 13 | 2.3.Analysis and use of assessment results | 0.456 | 1 | 4% |
| **14** | **4.5.Curricular emphasis on the subject** | **0.379** | **2** | **9%** |
| 15 | 4.4.Contribution of the activities to the achievement of the class objectives | 0.362 | 2 | 4% |
| 16 | 5.2.Quality of professional dialogue | 0.350 | 3 | 4% |
| 17 | 4.8.Encouragement | 0.275 | 2 | 4% |
| 18 | 5.1.Collaborative work suitability | 0.259 | 4 | 4% |
| 19 | 1.2.Relationship between activities and objectives | 0.240 | 1 | 4% |
| 20 | 1.1.Formation of learning objectives | 0.226 | 1 | 4% |

As it is shown in Table 4.19, the loadings that were situated in the second and third places when they were ordered from the highest loading to the smallest, corresponding to the two of the four factors with the highest theoretical weight in the

final score portfolio calculation. However, the other two indicators with the highest theoretical weight are in position ninth and fourteenth.

Table 4.20.
*Exploratory Confirmatory Factor Analysis loadings of the portfolio indicators from Teacher Evaluation 2017 with M3 results (N=11,216)*

| | Indicator | Factor loading | Factor | Weight |
|---|---|---|---|---|
| **1** | **4.7.Questions and activities** | **0.753** | **2** | **10%** |
| **2** | **4.6.Clear explanations** | **0.668** | **2** | **10%** |
| 3 | 4.3.Quality of the end of class | 0.624 | 2 | 5% |
| 4 | 2.3.Analysis and use of assessment results | 0.623 | 1 | 5% |
| **5** | **4.9. Student feedback** | **0.594** | **2** | **10%** |
| 6 | 4.2.Quality of the start of class | 0.593 | 2 | 5% |
| 7 | 4.1.Class environment | 0.555 | 2 | 5% |
| 8 | 2.2.Relationship between assessment and objectives | 0.538 | 1 | 5% |
| 9 | 3.1.Analyses based on students' characteristics | 0.491 | 2 | 5% |
| 10 | 1.2.Relationship between activities and objectives | 0.473 | 1 | 5% |
| 11 | 2.1. Evaluation and rubrics used for correction | 0.460 | 1 | 5% |
| 12 | 3.2.Use of error for learning | 0.457 | 2 | 5% |
| **13** | **4.5.Curricular emphasis on the subject** | **0.391** | **2** | **10%** |
| 14 | 4.4.Contribution of the activities to the achievement of the class objectives | 0.387 | **2** | 5% |
| 15 | 1.1.Formation of learning objectives | 0.356 | 1 | 5% |
| 16 | 4.8.Encouragement | 0.303 | 2 | 5% |

The results of teachers without an M3 score were similar. The loadings that were situated in the first and second places corresponded to the same two factors for the 3-factors solution and to those that have the highest theoretical weight in the final score portfolio calculation. Similarly, the other two indicators with the highest theoretical weight are in position fifth and thirteenth. The results, considering the factor loadings,

indicated that at least two of the four indicators with the highest weight in the final score (9% or 10%), presented the highest loadings in both models: 2-factor and 3-factor solutions.

## Weighted sum score

One advantage of the weighted sum score method is that items with the highest loading on the factor have the largest effect on the factor score (Distefano et al., 2009). Taking into consideration the weight of each indicator, which was outlined by the Chilean Teacher Evaluation System, the weighted score of each one of the portfolio indicators, modules, and portfolio was calculated.

The way in which the scores were calculated was the same used by the Evaluation System. That is, the result for each indicator category was assigned a score from 1 to 4, being unsatisfactory equal to 1, basic equal to 2, competent equal to 3, and outstanding equal to 4. Later, the total portfolio score was calculated as the average of the total indicators score, taking into consideration the respective weights proposed theoretically (4% or 5% for some indicators, and 9% or 10% for the others). Measures of central tendency for those weighted scores are presented in Table 4.21.

Table 4.21.
*Portfolio theoretical weighted scores for each Module*

| Sample | Module | N | M | SD | Min | Max |
|---|---|---|---|---|---|---|
| Teachers with M3 Score | 1 | 11,216 | 2.54 | 0.39 | 1 | 4 |
| | 2 | 11,212 | 2.39 | 0.35 | 1 | 3.77 |
| | 3 | 11,216 | 2.57 | 0.46 | 1 | 4 |
| | Total | 11,212 | 2.48 | 0.27 | 1 | 3.55 |
| Teachers without M3 Score | 1 | 10,216 | 2.42 | 0.43 | 1 | 3.86 |
| | 2 | 10,207 | 2.41 | 0.36 | 1 | 3.85 |
| | Total | 10,207 | 2.41 | 0.31 | 1 | 3.42 |
| Whole sample of teachers | Total | 21,428 | 2.45 | 0.35 | 1 | 3.86 |

In order to make the theoretical weighted score comparable, I transformed the latter to a standardized score. Measures of central tendency for those standardized weighted scores are presented in Table 4.22.

Table 4. 22.
*Portfolio theoretical weighted scores for each Module (standardized)*

| Sample | Module | N | M | SD | Min | Max |
|---|---|---|---|---|---|---|
| Teachers with M3 Score | 1 | 11,216 | 0.14 | 0.93 | -3.58 | 3.66 |
| | 2 | 11,212 | 0.00 | 1.00 | -3.97 | 3.91 |
| | 3 | 11,216 | 0.51 | 0.77 | -2.10 | 2.88 |
| | Total | 11,212 | 0.15 | 0.67 | -3.46 | 2.77 |
| Teachers without M3 Score | 1 | 10,216 | -0.15 | 1.05 | -3.58 | 3.32 |
| | 2 | 10,207 | 0.00 | 1 | -3.88 | 3.97 |
| | Total | 10,207 | -0.05 | 0.82 | -3.78 | 2.61 |

## Weighted Factor Scores.

The Weighted Factor Scores were calculated for each one of the three Modules for the teachers with the M3 score, and for the two Modules for teachers without the M3 score. Weighted factors were calculated, taking into consideration that items with higher loadings have a larger effect on the total factor score and vice versa. Therefore, before the addition, the factor loading of each item was multiplied to the scale score for each item. The advantage of this method is that it allows those indicators with the highest loadings on the factor to have the greatest effect on estimating the factor scores.

In order to calculate the weighted factor score for those teachers whose Module 3 was not included in their final score, the CFA analysis was calculated, considering the

two-factor structure described in aim 1 (one factor related to Module 1 and the other factor related to Module 2). Later, factor scores for each module were calculated.

For those teachers whose Module 3 was included in their final score, CFA analysis was calculated, considering three factors (one related to Module 1, the second related to Module 2, and the third related to Module 3). Measures of central tendency for those weighted factor scores are presented in Table 4.23.

*Table 4. 23. Portfolio factor scores for each Module*

| Sample | Module | N | M | SD | Min | Max |
|---|---|---|---|---|---|---|
| Teachers with M3 Score | 1 | 11,212 | .0005 | 0.84 | -2.46 | 2.66 |
| | 2 | 11,212 | .0002 | 0.85 | -3.67 | 3.21 |
| | 3 | 11,212 | .0003 | 0.74 | -2.43 | 1.95 |
| | Total | 11,212 | .0003 | 0.57 | -2.99 | 2.03 |
| Teachers without M3 Score | 1 | 10,207 | .0002 | 0.85 | -2.36 | 2.69 |
| | 2 | 10,207 | .0003 | 0.86 | -3.63 | 3.21 |
| | Total | 10,207 | .0003 | 0.67 | -3.07 | 1.98 |

Correlations between the standardized theoretical portfolio scores and the factor scores for the sample of teachers whose Module 3 was included in their final portfolio score are presented in Table 4.24. As can be seen in the table, the correlations between theoretical and factor scores were high for all of the three modules and for the total portfolio score, ranging between .90 and 0.95.

Table 4. 24.
*Correlations between theoretical weight and factor scores for teachers with M3*

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| 1. Module 1 score | 1.00 | | | | | | | |
| 2. Module 2 score | 0.16 | 1.00 | | | | | | |
| 3. Module 3 score | 0.39 | 0.21 | 1.00 | | | | | |
| 4. Module 1 factor score | **0.90** | 0.12 | 0.31 | 1.00 | | | | |
| 5. Module 2 factor score | 0.17 | **0.96** | 0.21 | 0.12 | 1.00 | | | |
| 6. Module 3 factor score | 0.34 | 0.19 | **0.93** | 0.27 | 0.19 | 1.00 | | |
| 7. Portfolio score | 0.72 | 0.73 | 0.65 | 0.62 | 0.71 | 0.59 | 1.00 | |
| 8. Portfolio factor score | 0.66 | 0.75 | 0.54 | 0.66 | 0.78 | 0.53 | **0.95** | 1.00 |

Correlations between the standardized theoretical portfolio scores and the factor scores for the sample of teachers whose Module 3 was not included in their final portfolio score are presented in table 4.25. As can be seen in the table, the correlations between theoretical and factor score were even higher for all the two modules and for the total portfolio score, ranging between .95 and 0.97.

Table 4. 25.
*Correlations between theoretical weight and factor scores for teachers without M3*

| | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| 1. Module 1 score | 1.00 | | | | | |
| 2. Module 2 score | 0.23 | 1.00 | | | | |
| 3. Module 1 factor score | **0.95** | 0.20 | 1.00 | | | |
| 4. Module 2 factor score | 0.24 | **0.97** | 0.20 | 1.00 | | |
| 5. Portfolio score | 0.77 | 0.80 | 0.71 | 0.78 | 1.00 | |
| 6. Portfolio factor score | 0.70 | 0.82 | 0.70 | 0.84 | **0.97** | 1.00 |

Paired *t*-test.

A dependent *t*-test was used to explore significant differences between the total portfolio standardized score, calculated using the Teacher Evaluation theoretical weights, and the portfolio total factor score. The results from the portfolio standardized weighted score for teachers with M3 results ($M$= 0.15, $SD$ = 0.67) and the total portfolio factor score ($M$ = 0.0003, $SD$ = 0.57), indicate that the mean differences between both ways to calculate the final portfolio score were different from 0, $t(11{,}211) = 73.85$, $p < .001$. Therefore, there were significant differences between the total weighted portfolio score and the total portfolio factor score for those teachers whose M3 was included in their final score.

With respect to the sample of teachers whose Module 3 was not taken into consideration for their final score, the results from the weighted standardized portfolio score ($M$= -0.05, $SD$ = 0.82) and the total portfolio factor score ($M$ = 0.0003, $SD$ = 0.67), indicate that the mean differences between both ways to calculate the final portfolio score were different from 0, $t(10{,}206) = -21.16$, $p < .001$. Therefore, there were significant differences between the total weighted portfolio score and the total portfolio factor score for those teachers whose M3 was not included in their final score.

Aim 4: Evaluate validity evidence that supports the interpretation and use of portfolio subscores.

Aim 4 evaluated if the portfolio subscores for each one of the modules has added value over the total score. In order to evaluate the use of module subscores, I calculated the value-added ratio (VAR; Feinberg and Wainer, 2014), which is a refined Haberman (2008) **proportional reduction of the mean squared error (PRMSE) method.**

The VAR is presented in a simple equation to approximate *PRMSEs* values:

$$\left(\frac{PRMSE_s}{PRMSE_x}\right) = VAR \approx 1.15 + 0.5 \times r_1 - 0.67 \times r_2$$

where $r_1$ is subscore reliability and $r_2$ is the disattenuated correlation between the subscore and the score remainder (composed of the remaining items on the test not included in the subset from which the subscore was computed (Feinberg & Jurich, 2017).

Table 4.26 shows the results from the value-added ratio for each module score in the portfolio for the sample of teachers whose Module 3 was included in their final score.

Table 4. 26.
*Subscore Value Added Ratio*

| | | **Subscore** | | |
|---|---|---|---|---|
| | **Statistics** | **Module 1** | **Module 2** | **Module 3** |
| Teachers with M3 Score | Subscore Reliability (*r*1) | 0.60 | 0.68 | 0.47 |
| | Remainder Subscore Reliability | 0.66 | 0.67 | 0.67 |
| | Raw Correlation | 0.32 | 0.22 | 0.36 |
| | Disattenuated Correlation (*r2*) | 0.51 | 0.32 | 0.64 |
| | **Equation** | **1.11** | **1.28** | **0.96** |

For the subscores to yield an added value over the total score, the results for the equations should be greater than one. As can be seen in Table 4.26, for the sample of teachers with Module 3 in their final score, the subscore report for Module 1 and Module 2 yielded added value over the total score (1.11 and 1.28, respectively). Conversely, for Module 3, no added value was found when reporting the score separately (0.96).

In the portfolio, Module 1 is made up of three different domains: planning, assessment, and reflection. Table 4.27 shows the results of the value-added ratio for Module 2, Module 3, and the three domains in Module 1. As can be seen in the table, from Module 1, only the assessment domain showed added value over the total score (1.16). Subscores for planning and reflection did not indicate an added value over the total portfolio score (0.82 and 0.98, respectively).

Table 4. 27.
*Subscore Value Added Ratio*

| | | Subscore | | | | |
|---|---|---|---|---|---|---|
| | **Statistics** | **Planning** | **Assessment** | **Reflection** | **Module 2** | **Module 3** |
| Teachers with M3 Score | Subscore Reliability ($r1$) | 0.16 | 0.57 | 0.44 | 0.68 | 0.47 |
| | Remainder Subscore Reliability | 0.71 | 0.68 | 0.69 | 0.67 | 0.67 |
| | Raw Correlation | 0.20 | 0.26 | 0.33 | 0.22 | 0.36 |
| | Disattenuated Correlation ($r2$) | 0.61 | 0.42 | 0.59 | 0.33 | 0.64 |
| | **Equation** | **0.82** | **1.16** | **0.98** | **1.28** | **0.96** |

For the group of teachers whose Module 3 was not included in their final portfolio score, Table 4.28 shows the results with respect to their value added subscores of the two modules.

Table 4. 28.
*Subscore Value Added Ratio*

| | | Subscore | |
|---|---|---|---|
| | **Statistics** | **Module 1** | **Module 2** |
| Teachers without M3 Score | Subscore Reliability (*r*1) | 0.67 | 0.71 |
| | Remainder Subscore Reliability | 0.71 | 0.67 |
| | Raw Correlation | 0.23 | 0.23 |
| | Disattenuated Correlation (*r*2) | 0.34 | 0.34 |
| | **Equation** | **1.27** | **1.29** |

As can be seen in the table, for the sample of teachers whose Module 3 was not included in their final score, both pedagogical material and video recorded class subscores showed had value over the total score (1.27 and 1.29). The results for the value added subscore, including the three domains that are part of Module 1, are presented in Table 4.29.

Table 4. 29.
*Subscore Value Added Ratio*

| | | Subscore | | | |
|---|---|---|---|---|---|
| | **Statistics** | **Planning** | **Assessment** | **Reflection** | **Module 2** |
| Teachers with M3 Score | Subscore Reliability (*r*1) | 0.29 | 0.62 | 0.53 | 0.71 |
| | Remainder Subscore Reliability | 0.72 | 0.70 | 0.70 | 0.67 |
| | Raw Correlation | 0.28 | 0.29 | 0.33 | 0.22 |
| | Disattenuated Correlation (*r*2) | 0.61 | 0.43 | 0.54 | 0.32 |
| | **Equation** | **0.89** | **1.18** | **1.06** | **1.28** |

As can be seen in the table, for the sample of teachers whose Module 3 was not included in their final score, the planning subscore did not yield value added over the total score (0.89). On the other hand, assessment, reflection, and Module 2 did yield value added over the total score (1.18 and 1.06).

In summary, for the evaluation of the possible subscores that could be reported in addition to the total portfolio score, the results indicated that for the teachers whose Module 3 was included in their final score, if the subscores for Module 1 (Pedagogical materials), and Module 2 (Video recorded class) are reported, that would yield an added value over the total score. Conversely, for Module 3 (Collaborative work), the results indicated that no added value was found if that score was reported separately. When a more detailed analysis is made considering the three domains separately that make up Module 1, the results indicated that only the domain of assessment showed added value over the total score.

Results for the teachers whose Module 3 was not included in their final score (only two modules analyzed), if the subscores for Module 1 (*Pedagogical materials*), and Module 2 (*Video recorded class*) were reported, both modules would yield an added value over the total score. With a more detailed analysis considering the three domains that make up Module 1, the results indicated that the domain of assessment and reflection showed added value over the total score.

# Chapter 5: Discussion

The main goal for the present research was to contribute to the body of research on the evidence of the validity of the Chilean Teacher Evaluation System, focusing specifically on the portfolio. As was previously mentioned, the Chilean Teacher Evaluation System takes into consideration different evaluation instruments. The present research focused specifically on one of them, the portfolio, because it could be considered the core of the evaluation system. It is also the only assessment shared by both systems that are currently used in Chile to evaluate teachers. Likewise, the portfolio is the most complex part of the evidence that the teachers have to submit for their evaluation process, and it contributes the most to the calculation of the overall score for the evaluation.

In order to contribute to the evidence of the validity of the Chilean Teacher Evaluation System, one important step was to clarify how the test scores would be interpreted and the purpose for which they would be used (Taut et al., 2012). For the Chilean Teacher Evaluation System, two broad purposes have been recognized by the stakeholders: formative and summative.

With the creation of the Chilean Teacher Professional Development System in 2016 the summative purpose of the system has grown in relevance. All the teachers that work in nationally funded schools will now be part of the Teaching Career. This implies that progressively from now to 2026, all the teachers that work in publicly funded schools

will be categorized in different levels of progression by demonstrating the skills and knowledge achieved. Therefore, in addition to the municipal teachers that have been evaluated since 2003, by 2026 around 108,000 teachers who work in charter and public early childhood schools are expected to be evaluated through a teacher performance portfolio and a standardized test of disciplinary and pedagogical knowledge in order to progress to the next Teaching Career level.

The Teacher Professional Development System goes hand in hand with a merit-based salary for teachers that could increase a teacher's salary as much as 30%. Thus, by 2026, 90% of the Chilean teachers will go through the evaluation process, which could highly impact their salary if they receive a favorable evaluation. The summative purpose for the Chilean Teacher Evaluation System, which is to support individual teacher rewards and sanctions based on the evaluation of teacher quality, is relieved in this new context. Therefore, the evidence of validity of the evaluation instruments become even more relevant for the Chilean Teacher Evaluation System, since they are intended to be used for high-stakes decisions.

Likewise, the formative purpose of the Chilean Teacher Evaluation System has not been shelved. The identification of teachers' strengths and weaknesses, in tandem with specific feedback reports about them have also been the purpose of the evaluation system. A more informative report has been developed, however for each instrument used in the evaluation, the results are reported by only one overall score for each instrument.

Taking these two purposes into consideration, four specific aims related to the assessment of portfolio internal structure were outlined. First, I assessed the structure of

the portfolio across two different subgroups of teachers via exploratory confirmatory factor analysis (ECFA). Second, I determined the portfolio structure invariance across different subgroups of teachers. Third, I compared the weight assigned to the portfolio indicators with empirical data using a weighted sum score method. And fourth, I evaluated if possible portfolio subscores showed added value over the total portfolio score.

From the four specific aims, there were four main findings with regard to portfolio validity from this dissertation. First, results from the ECFA indicate that a two-factor solution for the sample of teachers whose Module 3 was not part of their final score represented the two different aspects of the portfolio used to evaluate Chilean teachers. Conversely, the three-factor solution for those teachers whose Module 3 was part of their final score did not clearly represent the three theoretical dimensions proposed by the portfolio. The second finding was related to the portfolio structure invariance across teachers, depending on their school location and teaching level. The results showed a strong factorial invariance (invariant thresholds) across rural and urban teachers, and weak factorial invariance (invariant loadings) across the six groups of teaching levels. The third finding showed differences between the theoretical weight assigned to each portfolio indicator and the empirical weight. Finally, the fourth finding showed that Module 1 and Module 2 had added value over the total score for the portfolio. On the other hand, for those teachers who answered Module 3, that module did not show added value over the total score.

## Differences between teachers whose Module 3 was included or not included in their final score

Using ECFA for the whole sample of teachers evaluated in 2017, the theoretical model for the portfolio structure was validated. The results indicated that the three-factor structure solution presented a good model fit, and the loading showed the same structure as the proposed portfolio structure when the loading was larger than 0.3 (with the only exception of the indicator 4.8, with a loading close to 0.3). Similarly, for the teachers without Module 3, the structure with the two-factor solution was validated, with all of the indicators loading in their respective modules.

However, when I observed the results of the ECFA for the sample of teachers whose Module 3 was included in their final score, the theoretical three-factor solution was no longer supported. Instead, the results showed one factor that is made up of a combination of indicators from Module 1 and Module 3, suggesting that a unique factor was comprised of indicators from planning and reflection, but also collaborative work.

One possible explanation for this mixed factor could be that when I analyzed only those teachers for whom Module 3 was included in their final score, I included only those who improved their overall evaluation with such inclusion. Since Module 3 is not mandatory, one can infer that only highly motivated (abler) teachers are submitting their responses. Therefore, it is unclear whether the evaluation is rewarding collaborative work, or if it is picking up a self-selected sample of highly motivated teachers. To summarize, if by including Module 3, the sample only considers highly motivated, abler teachers, and they tend to be good at both collaborative work and at pedagogical practices, the differences between these two domains are expected to disappear, which is

what in fact happens. My results, despite not presenting concluding evidence that this is in fact the process at play, are suggestive of this as a plausible mechanism.

This last finding is important to take into consideration given that Module 3 was introduced into the evaluation process in 2017 after the incorporation of the law that created the Teaching Career. The purpose of including Module 3 was to enrich the evaluation of teacher quality, incorporating important aspects of professional educators beyond the classroom responsibilities, and related to the connection with the students outside of the classroom. This module includes the evaluation activities that are critical to preserving and enhancing the teaching profession. That addition was aligned with professional responsibilities, from the Good Teaching Framework (GTF; Domain D).

However, results from the present research came to show that the way in which Module 3 was included, as non-mandatory evidence to be incorporated, could imply a bias in the evaluation process. This is because this important aspect related to teacher quality as collaborative work is considered in the evaluation for only half of the teachers evaluated. However, for the other half of the teachers evaluated, their quality is evaluated only by taking into consideration the evidence of their pedagogical material presented and the observation from a recorded class. Therefore, the portfolio becomes an evaluation tool that does not differentiate between quality teachers, at least in the aspect of collaborative work. The portfolio then does not achieve its goal of evaluating teacher professional responsibilities, rather it suggests that only motivated, perhaps abler teachers, complete that aspect of the portfolio.

## Multigroup invariance

One of the objectives of teacher evaluation in the Chilean context is to be able to evaluate all teachers in the Chilean educational system using the same criteria as a way to preserve the objectivity of the evaluation. However, there is large heterogeneity in contexts and pedagogical strategies that teachers have to manage in order to cope with their classroom situation. For example, there are teachers who work in rural areas of the country, in schools that have the most impoverished population in Chile, and they face significant difficulties not found in urban settings. Moreover, a vast majority of those teachers need to adapt their pedagogical practices to multilevel classrooms; for instance, having to simultaneously work with students from first to fourth grade. In other cases, for the teacher evaluation process, it is possible to find teachers who teach children at completely different stages of their development. Until now, the early childhood teachers evaluated have been those that teach children at 4 and 5 years old. But, with the incorporation of more teachers into the evaluation process, by 2026 all early childhood teachers who receive public funding will be evaluated. Therefore, the same portfolio instrument will most likely be used to evaluate teachers who teach children from 6 months to 18 years old, as well as those teachers who teach adults. The present research used multigroup invariance in order to evaluate whether or not the indicators that comprise the portfolio operated equivalently across different teacher populations, such as school location and teaching level.

The results of measurement invariance considering urban and rural location indicated that the portfolio structure for teachers from both groups was invariant across those groups of teachers. Therefore, the structure of the portfolio composed of three

dimensions: pedagogical materials, video recording of a class, and collaborative work were the same for the teachers no matter where their school was located. Similarly, for rural and urban teachers, the factor loading across both groups was equivalent. Thus, the empirical weight for each indicator did not vary across teacher school location. Finally, when I looked for thresholds in each indicator from the four different possible performance categories, I observed that they were invariant across rural and urban teachers. Therefore, the thresholds of each indicator can be compared between urban and rural teachers.

In summary, for teachers depending on the rural or urban context where they work, the portfolio used in the Chilean Evaluation System operates equivalently across both populations of teachers, and the observed differences in the proportion of teachers evaluated in the four different possible categories of performance can be attributed to differences in the teacher quality construct evaluated.

Regarding the multigroup invariance for teachers depending on their teaching level, the results indicated that the different groups showed the same three-factor structure. Similarly, factor loading across the six groups of teachers was equivalent. Thus, the empirical weight for each indicator did not vary across teacher level. However, the indicators' threshold were not comparable across the six groups of teachers. The threshold between one category of performance to the other (unsatisfactory, basic, competent, and outstanding) on the portfolio indicators was not equivalent for the six teacher groups. Therefore, these results indicated that the performance categories for teachers depending on their teaching level would not be comparable to each other. The portfolio would not be captured through their performance levels evaluation rubrics

related to each one of the performance level of the portfolio, the different characteristics across teachers of different levels of teaching.

Taking this into consideration could lead to changes and adaptations to the portfolio evaluation rubrics, to better capture the differences from teachers from different levels of teaching. Recognizing previous efforts to standardized portfolio evaluations rubrics to different levels of teaching, these results suggest that more work is needed to better adapt the measurement to a wide array of contexts. The use of a "one-size-fits-all" type of portfolio evaluation rubric seems to be insufficient, especially considering its upcoming release as a single tool to measure teacher quality to almost all of the teachers in the country.

For instance, in order to make more adaptations to the rubric evaluation tools, a connection between the MIDE UC team, who carry out the technical work of the evaluation tools, and the educational institutions, in which these evaluation tools are applied, could be a way to understand the reality of different levels in order to contextualize the evaluation and adjust it to different teaching levels. Until now, specific work has been done between MIDE UC and the two largest providers of preschool education in Chile: JUNJI and Integra. A working agenda has been established to review the way in which the portfolio, as well as its evaluation rubrics, can be adapted to the reality of these two early childhood educational institutions. This has been important work since in those institutions there are educational levels that have not been previously evaluated, so the portfolio must be adapted to these realities.

## Weighted scores

As has been previously stated, the teacher evaluation portfolio is made up of 20 indicators. The final score is calculated by including the score for each one of these indicators (within a range of 1 to 4). Most of the indicators are weighted in the same way (4%), with the exception of four indicators that belong to Module 2: *curricular emphasis on the subject, clear explanations, questions and activities,* and *student feedback and use of assessment results*. The last four indicators have a higher weight (9%) compared to the remaining 16, basically because the teachers' skills that comprise them have been shown in the literature to have a strong correlation with the overall student learning.

In order to prove with empirical data whether or not those highly weighted variables for the Chilean teacher evaluation were, in fact, the same as suggested in theory, the weighted factor score for the whole portfolio was calculated. Using weighted factor score, the items with higher loadings have a larger effect on the total factor score.

The results regarding the factor loadings indicated that from the indicators that theoretically weigh the most, the indicators 4.7 (*questions and activities*), and 4.6 (*clear explanations*) also showed the higher loading. However, the other two indicators with the highest theoretical weight: 4.9 (*students' feedback*), and 4.5 (*curricular emphasis on the subject*), did not show high loading.

Additionally, with respect to the revision of the loading, weighted factor scores for the complete portfolio were calculated by averaging (weighted for the number of indicators) each one of the resulting factor scores. Later, the portfolio weighted factor score was compared to the theoretical portfolio score.

Significant differences appeared when comparing both portfolio scores: the factor weighted and theoretical scores. The results supported the idea that the weight that theoretically has been assigned to the portfolio does not necessarily match the empirical data. Therefore, more analysis with regard to the weight of each indicator had a great impact on the final portfolio score.

The results for the comparison between theoretical and empirical weights indicated that the theoretical weight is partially validated. Two of the indicators show coincidence, but the other two do not. This is important to review because by indicating that certain indicators weigh more than others, the emphasis is being placed on certain characteristics related to teaching quality that are considered more relevant to determine well-evaluated teachers than those that do not. If this emphasis does not coincide with the data information, a revision of the weights is necessary.

### Subscores' added value

The portfolio used in the Chilean teacher evaluation system has at least three different subsections based on content areas: pedagogical material, video recorded class, and collaborative work. Within the pedagogical material area, three possible subsections have been described: planning, assessment, and reflection. Until now, The Ministry of Education has only reported a total score as the final portfolio evaluation. However, the teachers' increasing demands for subscore information (to better understand the specific areas in which they could improve their weaknesses in their evaluation) led me to focus on this particular area.

Given the importance of subscore reporting, the quality of the subscores must be assessed in order to avoid inaccurate information. Using the value-added ratio (VAR), I

assessed if each one of the modules included in the portfolio yielded an added value over the total score. Results indicated that reporting for Module 1 and Module 2 yielded added value over the total score. On the other hand, reporting for Module 3 score did not show added value over the total score. I also assessed the added value within the three different subdomains in Module 3. The results indicated that only the results for assessment showed added value over the total score.

The results for added value were positive, as there is evidence that supports the idea of reporting information from at least two of the three modules evaluated by the teacher evaluation system. Likewise, information for the assessment domain was also reported. Being able to report extra information about the results is valuable for the system in general and the teachers; more information can help teachers to focus on their weaknesses. Also, more information can orient the system for future teacher development plans, allocating the resources in a more efficient and effective way. To illustrate, when two different teachers obtain the same overall score, they are subject to the same professional development plan to improve their skills for the next evaluation process, regardless of their respective specificities and issues. Being able to pinpoint the specific areas of improvement in each case represents an opportunity for the Chilean government and the Ministry of Education to reduce public spending and to improve and accelerate their results, in terms of teacher quality. A strategy like this one can cost-effective and can provide better results.

In summary, results for the present dissertation indicated that, in general, the portfolio used to evaluate the teachers for the Chilean Teacher Evaluation System is a valid evaluation instrument for the two purposes for which it was created: summative and

formative. With respect to the summative purpose, the evidence validates a portfolio that evaluates the quality of teachers in three main constructs: pedagogical practices, observed video class, and collaborative work for the complete sample of teachers evaluated, and for two main constructs: pedagogical practices and observed video class, for the teachers without Module 3 evidence. However, for the sample of teachers whose score in Module 3 was considered for their final portfolio score, the evidence did not confirm the theoretical portfolio structure.

Evidence presented for portfolio validity also related to the summative purpose was about the portfolio equivalence across different populations such as school location (e.g., teachers from urban schools and teachers from rural schools), or teaching levels (e.g., preschool teachers and high school teachers). This evidence pretended to answer the question of the portfolio appropriateness distinguishing quality teachers regardless of the context or educational level in which the teachers teach. Results showed portfolio validity with respect to the invariance of the structure, factor loadings, and threshold between the four possible performance levels, for the teachers working in rural and urban contexts. However, with respect to the teaching levels, the portfolio was only valid with respect to the invariance at the structure and factor loadings levels. These results indicated that at least for rural and urban teachers, the portfolio correctly distinguishes the differences in the teacher quality construct evaluated.

The last aim related to evidence of validity considering the summative purpose of the portfolio was the comparison between theoretical and empirical weights of each portfolio indicator. With the empirical evidence, the portfolio weights were partially

validated since at least two of the four highest weighted indicators coincided with the empirical information.

With respect to the validity of the portfolio with regards to their formative purpose, more information reported is better. The results for this dissertation indicated that at least for two of the three modules, the inclusion of more detailed information of possible portfolio subscores would be valid.

Overall, the portfolio used in the Chilean Teacher Evaluation System is a good evaluation tool of teacher quality, of which it is necessary to review the incorporation of collaborative work evaluation for all of the teachers evaluated, the specific rubric evaluations for teachers from different teaching levels, and the theoretical weights. However, the portfolio is an instrument that fulfills its summative and formative purpose, signaling the quality of teachers in Chile, as well as delivering timely feedback. This instrument could be an example to some of the beginning evaluation systems in other Latin American countries, who can take this instrument as a starting point in creating their own teacher evaluation systems.

## Limitations

One limitation of the present study is that the evaluation results did not include information from charter school teachers. Since the implementation of the Teacher Professional Development System, which has included the charter school teachers in the evaluation process from 2016 on in a trial setting, the first results from those evaluations were from the 2018 cohort (results that are still not available for research purposes). A second limitation of the present dissertation is the fact that I only focused on one type of validity evidence described by the Standards, which is internal structural validity. While

this is an important source of evidence regarding instrument validity, the Standards describe other sources equally important in the instrument validation process. Nevertheless, important information about structural validity was produced in this research that can be used for the system feedback. Third, the current study focused on only one of the four instruments used in the Chilean Teacher Evaluation and one of the two instruments that are being used as part of the Chilean Teacher Professional Development System. More information from the other instruments can be used in order to produce evidence for structural validity, but the focus for the present dissertation was on the portfolio instrument since it is the core instrument between both evaluation systems that are currently used in Chile. Finally, the present research did not include information from technical teachers since their portfolio contains different indicators than the ones used to evaluate the other teachers. That sample of teachers removed represented 9.4% of the complete sample of teachers evaluated in 2017.

## Future Research

Future research should concentrate on the inclusion of the other four types of sources of validity evidence defined by the Standards, such as: evidence based on test content, evidence based on the response process, evidence based on the connection between variables, and evidence of the consequences of testing.

Likewise, more research about Module 3 is necessary in order to understand the reasons behind the high non-response rate for this module, as well as the consequences for the final teacher score of possibly making this module mandatory in the future. Since Module 3 showed the lowest overall results when I analyzed the whole sample of teachers, the implications of making it mandatory suggest that individual teacher scores

will go down. These potential lower scores would impact the whole teacher evaluation system and could have potentially negative consequences, as teachers may feel overwhelmed by the process.

Since the ascension in the Teaching Career System depends on the score of the standardized test of disciplinary and pedagogical knowledge, future research should include the information from those results as evidence validity based on the connection between variables. It will also be important to research the type of relationship between teacher practices (portfolio) and disciplinary knowledge (test), taking into consideration the possibilities of threshold or plateaus in this association. For instance, this research could be important for focusing on the implementation of professional developmental plans until the teachers reach a minimum threshold of knowledge to have better teaching practices.

Finally, with the inclusion of charter school teachers, future research related to the invariance with respect to municipal teachers could be important in order to evaluate how a unique instrument operates equivalently across these different populations of teachers.

References

Aaronson, D., Barrow, L., & Sander, W. (2007). Teachers and Student Achievement in the Chicago Public High Schools. *Journal of Labor Economics*, *25*(1), 95–135. https://doi.org/10.1086/508733

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

Alvarado, M., Cabezas, G., Falck, D., & Ortega, M. E. (2011). *La Evaluación Docente y sus instrumentos: Discriminación del desempeño docente y asociación con los resultados de los estudiantes*. PNUD, Centro de Estudios Mineduc.

Anastasi, A. (1986). Evolving Concepts of Test Validation. *Annual Review of Psychology*, *37*(1), 1–16. https://doi.org/10.1146/annurev.ps.37.020186.000245

Angoff, W. H. (1988). Validity: An envolving concept. In *Test Validity* (H. Wainer&H.I. Braun, pp. 19–32). Lawrence Erlbaum Associates.

Aragon, S. (2018). *Teacher evaluations: What is the issue and why does it matter? Policy snapshot*. Education Comission of the States. https://www.ecs.org/wp-content/uploads/Teacher_Evaluations.pdf

Asparouhov, T., & Muthén, B. (2009). Exploratory Structural Equation Modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *16*(3), 397–438. https://doi.org/10.1080/10705510903008204

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

Beasley, T. M., & Schumacker, R. E. (1995). Multiple Regression Approach to Analyzing Contingency Tables: Post Hoc and Planned Comparison Procedures. *The Journal of Experimental Education*, *64*(1), 79–93. https://doi.org/10.1080/00220973.1995.9943797

Bentler, P. M. (1995). *EQS structural equations program manual*. Multivariate Software, Inc.

Bill & Melinda Gates Foundation. (2012). *Gathering Feedback for Teaching. Combining High-Quality Observations with Student Surveys and Achievement Gains*.

Bollen, K. A. (1989). *Structural equations with latent variables*. Wiley.

Bonifaz, R. (2011). Origen de la Evaluación Docente y su conexión con las políticas públicas en educación. In *La evaluación docente en Chile* (J. Manzi, R. González, Y. Sun, pp. 13–32). Mide UC.

Bowen, N. K., & Masa, R. D. (2015). Conducting Measurement Invariance Tests with Ordinal Data: A Guide for Social Work Researchers. *Journal of the Society for Social Work and Research*, *6*(2), 229–249. https://doi.org/10.1086/681607

Brown, T. A. (2015). *Confirmatory factor analysis for applied research* (Second edition). The Guilford Press.

Bruns, B., & Luque, J. (2014). *Great Teachers: How to Raise Student Learning in Latin America and the Caribbean*. The World Bank. https://doi.org/10.1596/978-1-4648-0151-8

Bryk, A., Harding, H., & Greenberg, S. (2012). Contextual Influences on Inquiries into Effective Teaching and Their Implications for Improving Student Learning.

*Harvard Educational Review*, *82*(1), 83–106. https://doi.org/10.17763/haer.82.1.k58q7660444q1210

Byrne, B. M. (2008). Testing for Multigroups Equivalence of a Measuring Instrument: A Walk Through the Process. *Psicothema*, *20*(4), 872–882.

Byrne, B. M. (2012). *Structural equation modeling with Mplus: Basic concepts, applications, and programming*. Routledge Academic.

Carbaugh, B., Marzano, R., & Toth, M. (2017). *The Marzano Focused Teacher Evaluation Model*. LearningSciences Marzano Center.

Carmines, E. G., & Zeller, R. A. (1979). *Reliability and validity assessment*. Sage Publications.

Castillo-Miranda, S. del R., Castro, G. W., & Hidalgo-Standen, C. (2017). La Evaluación del Desempeño Docente desde la perspectiva de profesores de educación rural. *Educación y Educadores*, *20*(3), 364–381. https://doi.org/10.5294/edu.2017.20.3.2

Centro de Estudios. (2017). *Análisis de la Reforma Educacional en base a los principales indicadores del Education at a Glance 2017*. Ministerio de Educación de Chile.

Centro de Medición, MIDE UC. (2019). *Grupo de Estudios Carrera Docente. Informaciones Generales*.

Chen, F. F. (2007). Sensitivity of Goodness of Fit Indexes to Lack of Measurement Invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *14*(3), 464–504. https://doi.org/10.1080/10705510701301834

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2011). *The Long-Term Impacts of Teachers: Teacher Value-Added and Student Outcomes in Adulthood*. NBER WORKING PAPER SERIES.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating Goodness-of-Fit Indexes for Testing Measurement Invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *9*(2), 233–255. https://doi.org/10.1207/S15328007SEM0902_5

Cleaver, S., Detrich, R., & States, J. (2018). *Overview of Teacher Formal Evaluation*. The Wing Institute. https://www.winginstitute.org/teacher-evaluation-formal.

Clinton, J., Anderson, M., Dawson, G., Dawson, A., Bolton, S., Mason, R., Australia, Department of Education and Training, University of Melbourne, & Centre for Program Evaluation. (2017). *Teacher effectiveness systems, frameworks and measures: A review*. http://nla.gov.au/nla.obj-397261266

Colegio de Profesores de Chile A.G. (2016). *La evaluación docente en el mundo rural*. Www.Colegiodeprofesores.Cl. http://www.colegiodeprofesores.cl/la-evaluacion-docente-en-el-mundo-rural/

Cortes, F., & Lagos, M. J. (2011). Consecuencias de las Evaluacion Docente. In *La evaluación docente en Chile* (J. Manzi, R. González, Y. Sun, pp. 137–154). Mide UC.

Cortez-Ochoa, A., Thomas, S., Tikly, L., & Doyle, H. (2018). *Scan of International Approaches to Teacher Assessment*. Univertisy of Briston, School of Education.

Crocker, L. (1997). Assessing Content Representativeness of Performance Assessment Exercises. *Applied Measurement in Education*, *10*(1), 83–95. https://doi.org/10.1207/s15324818ame1001_5

Cronbach, L. J. (1988). Five perspectives on validation argument. In *Test Validity* (H. Wainer, H. Braun, pp. 3–18). L. Erlbaum Associates.

Cruz-Aguayo, Y., Hincapé, D., & Rodríguez, C. (2020). *Profesores a prueba: Claves para una evaluación docente exitosa*. Inter-American Development Bank. https://doi.org/10.18235/0002149

Danielson, C. (2007). *Enhancing professional practice: A framework for teaching* (2nd ed). Association for Supervision and Curriculum Development.

Darling-Hammond, L. (2000). Teacher Quality and Student Achievement. *Education Policy Analysis Archives*, *8*, 1. https://doi.org/10.14507/epaa.v8n1.2000

Darling-Hammond, L. (2010). *Evaluating teacher effectiveness: How teacher performance assessments can measure and improve teaching.* (Center for American Progress).

Department for Education. (2019). *Teacher appraisal and capability. A model policy for schools*. Department for Education, UK.

DeVellis, R. F. (2017). *Scale development: Theory and applications* (Fourth edition). SAGE.

Dimitrov, D. M. (2010). Testing for Factorial Invariance in the Context of Construct Validation. *Measurement and Evaluation in Counseling and Development*, *43*(2), 121–149. https://doi.org/10.1177/0748175610373459

Distefano, C., Zhy, M., & Mindrila, D. (2009). Understanding and Using Factor Scores: Considerations for the Applied Researcher. *Practical Assessment, Research & Evaluation*, *14*(20), 1–11.

Docentemas. (2019). *Que es la Evaluacion Docente*.

Docentemas. (2020). *Corrección de Portafolios*. Docentemas.Cl.

Donker, A. S., de Boer, H., Kostons, D., Dignath van Ewijk, C. C., & van der Werf, M. P. C. (2014). Effectiveness of learning strategy instruction on academic performance: A meta-analysis. *Educational Research Review*, *11*, 1–26. https://doi.org/10.1016/j.edurev.2013.11.002

Drasgow, F., & Kanfer, R. (1985). Equivalence of psychological measurement in heterogeneous populations. *Journal of Applied Psychology*, *70*(4), 662–680. https://doi.org/10.1037/0021-9010.70.4.662

Educational Testing Service. (2018). *PPAT® Assessment Candidate and Educator Handbook*. Educational Testing Service.

Enders, C. K. (2010). *Applied missing data analysis*. Guilford Press.

Everson, K. C., Feinauer, E., & Sudweeks, R. (2013). Rethinking Teacher Evaluation: A Conversation about Statistical Inferences and Value-Added Models. *Harvard Educational Review*, *83*(2), 349–370. https://doi.org/10.17763/haer.83.2.m32hk8q851u752h0

Feinberg, R. A., & Jurich, D. P. (2017). Guidelines for Interpreting and Reporting Subscores. *Educational Measurement: Issues and Practice*, *36*(1), 5–13. https://doi.org/10.1111/emip.12142

Feinberg, R. A., & Wainer, H. (2014). A Simple Equation to Predict a Subscore's Value. *Educational Measurement: Issues and Practice*, *33*(3), 55–56. https://doi.org/10.1111/emip.12035

Feinberg, R. A., & Wainer, H. (2015). How Much Is Enough? A Reply to Sinharay, Haberman, and Boughton. *Educational Measurement: Issues and Practice*, *34*(3), 9–9. https://doi.org/10.1111/emip.12081

Ferguson, R. F., & Danielson, C. (2015). How Framework for Teaching and Tripod 7Cs Evidence Distinguish Key Components of Effective Teaching. In T. J. Kane, K. A. Kerr, & R. C. Pianta (Eds.), *Designing Teacher Evaluationig Systems* (pp. 98–143). John Wiley & Sons, Inc. https://doi.org/10.1002/9781119210856.ch4

Figueroa, M., García, S., Maldonado, D., Rodriguez, C., Saavedra, A. M., & Vargas, G. (2018). *La profesión docente en Colombia: Normatividad, formación, selección y evaluación*. Universidad de los Andes - Escuela de Gobierno Alberto Lleras Camargo.

Flotts, M. P., & Abarzua, A. (2011). El modelo de evaluación y los instrumentos. In *La evaluación docente en Chile* (J. Manzi, R. González, Y. Sun, pp. 35–61). Mide UC.

Gerritsen, S., Plug, E., & Webbink, D. (2017). Teacher Quality and Student Achievement: Evidence from a Sample of Dutch Twins: Teacher Quality and Student Achievement. *Journal of Applied Econometrics*, *32*(3), 643–660. https://doi.org/10.1002/jae.2539

Goe, L. (2007). *The Link between Teacher Quality and Student Outcomes: A Research Synthesis* (p. 72). National Comprehensive Center for Teacher Quality.

Goepel, J. (2012). Upholding public trust: An examination of teacher professionalism and the use of Teachers' Standards in England. *Teacher Development*, *16*(4), 489–505. https://doi.org/10.1080/13664530.2012.729784

Goldhaber, D. (2002). The mystery of good teaching. *Education Next*, *2*, 50–55.

Haberman, S. J. (2008). When Can Subscores Have Value? *Journal of Educational and Behavioral Statistics*, *33*(2), 204–229. https://doi.org/10.3102/1076998607302636

Hakel, M. D., Koenig, J. A., Elliott, S. W., & National Board for Professional Teaching Standards (U.S.) (Eds.). (2008). *Assessing accomplished teaching: Advanced-level certification programs: Committee on Evaluation of Teacher Certification by the National Board for Professional Teaching Standards*. National Academies Press.

Haladyna, T. M., & Kramer, G. A. (2004). The Validity of Subscores for a Credentialing Test. *Evaluation & the Health Professions*, *27*(4), 349–368. https://doi.org/10.1177/0163278704270010

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Sage Publications.

Harris, D. N., & Sass, T. R. (2011). Teacher training, teacher quality and student achievement. *Journal of Public Economics*, *95*(7–8), 798–812. https://doi.org/10.1016/j.jpubeco.2010.11.009

Hattie, J. (2008). Chapter 4 Validating the specification of standards for teaching: Applications to the National Board for Professional Teaching Standards' Assessments. In *Advances in Program Evaluation* (Vol. 11, pp. 93–111). Emerald (MCB UP ). https://doi.org/10.1016/S1474-7863(07)11004-8

Hattie, J. (2012). *Visible learning for teachers: Maximizing impact on learning*. Routledge.

Hattie, J., & Timperley, H. (2007). The Power of Feedback. *Review of Educational Research*, *77*(1), 81–112. https://doi.org/10.3102/003465430298487

Heck, R. H. (2009). Teacher effectiveness and student achievement: Investigating a multilevel cross-classified model. *Journal of Educational Administration*, *47*(2), 227–249. https://doi.org/10.1108/09578230910941066

Hirschfeld, G., & Brachel, R. (2014). Improving Multiple-Group confirmatory factor analysis in R – A tutorial in measurement invariance with continuous and ordinal indicators. *Practical Assessment, Research, and Evaluation*, *19*, Article 7.

Hu, L., & Bentler, P. (1995). Evaluating model fit. In *Structural equation modeling: Concepts, issues, and applications* (R.H. Hoyle, pp. 76–99). Sage Publications, Inc.

Huber, S. G., & Skedsmo, G. (2016). Editorial: Data Use—a Key to Improve Teaching and Learning? *Educational Assessment, Evaluation and Accountability*, *28*(1), 1–3. https://doi.org/10.1007/s11092-016-9239-8

Isoré, M. (2009). *Teacher Evaluation: Current Practices in OECD Countries and a Literature Review* (OECD Education Working Papers No. 23). https://doi.org/10.1787/223283631428

Jaeger, R. M. (1998). Evaluating the Psychometric Qualities of the National Board for Professional Teaching Standards' Assessments: A Methodological Accounting. *Journal of Personnel Evaluation in Education*, *12*(2), 189–210. https://doi.org/10.1023/A:1008085128230

Jordan, H. R., Mendro, R. L., & Weersinghe, D. (1997). *Teacher effects on longitudinal student achievement: A preliminary report on research on teacher effectiveness*. National Evaluation Institute, Indianapolis, IN. Kalamazoo, MI.

Kane, Michael T. (2013). Validating the Interpretations and Uses of Test Scores: Validating the Interpretations and Uses of Test Scores. *Journal of Educational Measurement*, *50*(1), 1–73. https://doi.org/10.1111/jedm.12000

Kane, M.T. (2006). Validation. In *Educational measurement* (pp. 17–64). Praege.

Kane, T. J., Kerr, K. A., Pianta, R. C., & Measures of Effective Teaching Project (Eds.). (2014). *Designing teacher evaluation systems: New guidance from the measures of effective teaching project* (First edition). Jossey-Bass, a Wiley Brand.

Kang, H. (2013). The prevention and handling of the missing data. *Korean Journal of Anesthesiology*, *64*(5), 402. https://doi.org/10.4097/kjae.2013.64.5.402

Koretz, D. M. (2008). *Measuring up: What educational testing really tells us*. Harvard University Press.

Lazarev, V., Newman, D., & Sharp, A. (2014). *Properties of the multiple measures in Arizona's teacher evaluation model*. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory West.

Little, T. D. (1997). Mean and Covariance Structures (MACS) Analyses of Cross-Cultural Data: Practical and Theoretical Issues. *Multivariate Behavioral Research*, *32*(1), 53–76. https://doi.org/10.1207/s15327906mbr3201_3

Looney, J. (2011). Developing High-Quality Teachers: Teacher evaluation for improvement: European Journal of Education, Part I. *European Journal of Education*, *46*(4), 440–455. https://doi.org/10.1111/j.1465-3435.2011.01492.x

Manzi, J., & Jiménez, D. (2017, May). *¿Qué evidencia respalda la validez de la evaluación docente en Chile?* [Seminario]. Seminario MIDE UC, Santiago, Chile.

Martínez, J. F., Borko, H., & Stecher, B. M. (2012). Measuring instructional practice in science using classroom artifacts: Lessons learned from two validation studies. *Journal of Research in Science Teaching*, *49*(1), 38–67. https://doi.org/10.1002/tea.20447

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

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, *50*(9), 741–749. https://doi.org/10.1037/0003-066X.50.9.741

Milanowski, A. (2011). *Validity Research on Teacher Evaluation Systems Based on the Framework for Teaching*. American Education Research Association annual meeting, New Orleans.

Miller, M. D., Linn, R. L., Gronlund, N. E., & Linn, R. L. (2009). *Measurement and assessment in teaching* (10th ed). Merrill/Pearson.

Ministerio de Educacion. (2003). *Estándares para la Formación Inicial Docente*. MINEDUC.

Ministerio de Educacion, Centro de Estudios, & Unidad de Estadisticas. (2018). *Estadisticas de la Educacion 2017* (MINEDUC).

Ministerio de Educacion de Chile. (2008). *Marco para la Buena Enseñanza [Good Teaching Framework]*. MINEDUC.

Ministerio de Educación Perú. (2020). *Evaluación del Desempeño Docente*. Evaluaciondocente. http://evaluaciondocente.perueduca.pe/desempenoinicialtramo2/mas_informacion/

OECD. (2009). *Teacher Evaluation. A Conceptual Framework and examples of Country Practices*. OECD-Mexico Workshop Towards a Teacher Evaluation Framework in Mexico: International Practices, Criteria and Mechanisms, Mexico City.

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

OECD. (2017). *Education in Chile*. OECD. https://doi.org/10.1787/9789264284425-en

Rios, J., & Wells, C. (2014). Validity evidence based on internal structure. *Psicothema*, *26.1*, 108–116. https://doi.org/10.7334/psicothema2013.260

Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, Schools, and Academic Achievement. *Econometrica*, *73*(2), 417–458. https://doi.org/10.1111/j.1468-0262.2005.00584.x

Rockoff, J. E. (2004). The Impact of Individual Teachers on Student Achievement: Evidence from Panel Data. *American Economic Review*, *94*(2), 247–252. https://doi.org/10.1257/0002828041302244

Rockoff, J. E., & Speroni, C. (2011). Subjective and objective evaluations of teacher effectiveness: Evidence from New York City. *Labour Economics*, *18*(5), 687–696. https://doi.org/10.1016/j.labeco.2011.02.004

Ross, E., & Walsh, K. (2019). *State of the States 2019: Teacher & Principal Evaluation Policy*. National Council on Teacher Quality.

Ruffinelli, A. (2016). Law on Teacher Professional Development: From systematic precariousness to the achievements, the advances, and the remaining challenges for professionalization. *Estudios Pedagogicos*, *XLII*(4), 261–279.

Sahlberg, P. (2011). The Fourth Way of Finland. *Journal of Educational Change*, *12*(2), 173–185. https://doi.org/10.1007/s10833-011-9157-y

Sanders, W. L., & Rivers, J. C. (1996). *Cumulative and residual effects of teachers on future student academic achievement.* Research Progress Report. Knoxville: University of Tennessee Value-Added Research and Assessment Center.

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

Santiago, P., & OECD (Eds.). (2013). *Teacher evaluation in Chile 2013*. OECD Publishing.

Sartain, L., Brown, E. R., Stoelinga, S. R., & Consortium on Chicago School Research. (2011). *Rethinking Teacher Evaluation in Chicago: Lessons Learned from ClassroomObservations, Principal-Teacher Conferences, and DistrictImplementation. Research Report.*

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, *7*(2), 147–177. https://doi.org/10.1037/1082-989X.7.2.147

Schmelkes, S. (2015). Assessment of teacher performance – State of affairs. In *Critical issues for formulating new teacher policies in Latin America and the Caribbean: The current debate*. OREALC/UNESCO.

Sevilla, M. P. (2011). *Educación Técnica Profesional en Chile. Antecedentes y claves de diagnóstico*. Ministerio de Educación de Chile.

Sinharay, S. (2010). How Often Do Subscores Have Added Value? Results from Operational and Simulated Data: Subscores. *Journal of Educational Measurement*, *47*(2), 150–174. https://doi.org/10.1111/j.1745-3984.2010.00106.x

Sinharay, S., Haberman, S., & Boughton, K. (2015). Too Simple to Be Useful: A Comment on Feinberg and Wainer (2014). *Educational Measurement: Issues and Practice*, *34*(3), 6–8. https://doi.org/10.1111/emip.12080

Sinharay, S., & Haberman, S. J. (2008). *Reporting Subscores: A Survey*. ETS.

Skedsmo, G., & Huber, S. G. (2018). Teacher evaluation: The need for valid measures and increased teacher involvement. *Educational Assessment, Evaluation and Accountability*, *30*(1), 1–5. https://doi.org/10.1007/s11092-018-9273-9

Stanford Center for Assessment, Learning and Equity. (2013). *2013 edTPA Field Test: Summary Report*. Stanford Center for Assessment, Learning and Equity.

Steiner, L. (2010). *Using competency-based evaluation to drive teacher excellence: Lessons from Singapore.* Public Impact; https://schoolturnaroundsupport.org/sites/default/files/resources/usingcompetency

basedeval.pdf. https://schoolturnaroundsupport.org/sites/default/files/resources/usingcompetencybasedeval.pdf

Stronge, J. (2010). *Teacher Performance Evalaution System*. Association of America Schools in South America (AASSA).

Stronge, J. (2012). *Teacher Effectiveness Performance Evaluation System*. Educational Consulting, LLC.

Sun, Y. (2018, April 10). *La Evaluación Docente en Chile y su evoluación* [Conference]. III Congreso Latinoamericano de Medición y Evaluación Educacional, Montevideo, Uruguay.

Sun, Y., Calderon, P., Valerio, N., & Torres, P. (2011). La implementacion de la Evaluacion Docente. In *La evaluación docente en Chile* (J. Manzi, R. González, Y. Sun, pp. 63–89). Mide UC.

Tarhan, H., Karaman, A. C., Kemppinen, L., & Aerila, J.-A. (2019). Understanding Teacher Evaluation in Finland: A Professional Development Framework. *Australian Journal of Teacher Education*, *44*(4), 33–50. https://doi.org/10.14221/ajte.2018v44n4.3

Tate, R. (2002). Test Dimensionality. In *Large-scale assessment programs for all students: Validity, technical adequacy, and implementation* (Lawrence Erlbaum, pp. 181–211). J. Tindal&T.M. Haladyna.

Taut, S., Santelices, M. V., & Manzi, J. (2011). Estudios de validez de la Evaluacion Docente. In *La evaluación docente en Chile* (J. Manzi, R. González, Y. Sun, pp. 155–175). Mide UC.

Taut, S., Santelices, M. V., & Stecher, B. (2012). Validation of a National Teacher Assessment and Improvement System. *Educational Assessment*, *17*(4), 163–199. https://doi.org/10.1080/10627197.2012.735913

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*. https://doi.org/10.14507/epaa.v22n71.2014

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

The National Center on Education and the Economy. (2016). *Singapore: A Teaching Model for the 21st Century* (Empowered Educators). The National Center on Education and the Economy.

Toch, T. (2008). Fixing Teacher Evaluation. *Educational Leadership*, *66*(2), 32–37.

Tornero, B., & Taut, S. (2010). A mandatory, high-stakes National Teacher Evaluation System: Perceptions and attributions of teachers who actively refuse to participate. *Studies in Educational Evaluation*, *36*(4), 132–142. https://doi.org/10.1016/j.stueduc.2011.02.002

Tripod Education Partner. (2014). *The Tripod 7Cs of Effective Teaching An Overview*.

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

Vaillant, D. (2008). Algunos marcos referenciales para el evaluacion del desempeño docente en América Latina. *Revista Iberoamericana de Evaluación Educativa*, *1*(2), 7–22.

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

Vartuli, S. (1999). How early childhood teacher beliefs vary across grade level. *Early Childhood Research Quarterly*, *14*(4), 489–514. https://doi.org/10.1016/S0885-2006(99)00026-5

Vázquez Cruz, M. del Á., Cordero Arroyo, G., & Leyva Barajas, Y. E. (2014). Análisis comparativo de criterios de desempeño profesional para la enseñanza en cuatro países de América / Comparative analysis of professional performance criteria for teaching in four countries of America. *Actualidades Investigativas En Educación*, *14*(3). https://doi.org/10.15517/aie.v14i3.16086

Villarroel, G. (2003). El profesor rural de Chiloé. *Revista Digital ERural, Educación, Cultura y Desarrollo Rural.*

Wright, S. P., Horn, S., & Sanders, W. L. (1997). Teacher and Classroom Context Effects on Student Achievement: Implications for Teacher Evaluation. *Journal of Personnel Evaluation in Education*, *11*(1), 57–67. https://doi.org/10.1023/A:1007999204543

Zelda, B., & Sánchez, F. (2017). *Does Better Teacher Selection Lead to better Students? Evidence from a Large Scale Reform in Colombia*. Documentos CEDE 015350, Universidad de los Andes.