Discussion Paper Series

IZA DP No. 18505

March 2026

When Teacher Preparation Programs Look Alike: Variability, Accountability, and the Limits of Program Differentiation

Rosario Rivero
Universidad Diego Portales and Talento al Aula

Rafael Sánchez
CUNEF Universidad and IZA@LISER

Edgar Valencia
Pontificia Universidad Católica de Chile

María Eugenia Rojas
Pontificia Universidad Católica de Valparaíso and CIAE-Universidad de Chile

When Teacher Preparation Programs Look Alike: Variability, Accountability, and the Limits of Program Differentiation*

Abstract

Research on teacher preparation programs (TPPs) continues to debate the extent to which program quality meaningfully shapes teacher effectiveness. While early evidence from high-income contexts documented nontrivial differences across programs, more recent studies suggest that most variation in teacher effectiveness occurs within rather than between programs. Using national administrative data from Chile and a three-level value-added model linking students, teachers, and preparation programs, this study estimates the share of variance in student achievement attributable to TPPs. Focusing on novice mathematics teachers linked to the only cohort with available lagged test scores (students assessed in 2015 and 2017), we find that TPPs account for approximately 1% of the total variance in student outcomes, while substantially more variation occurs within programs across teachers. Rather than interpreting this limited differentiation as evidence of uniformly strong preparation, we discuss how these patterns are consistent with institutional convergence and bounded instructional learning in a highly regulated system. The study contributes new empirical evidence from Latin America and offers a theoretically grounded interpretation of how accountability regimes may shape the distribution — rather than the overall level — of teacher effectiveness across preparation programs.

JEL classification

I20, I28, I29

Keywords

teacher preparation programs, teacher education, teacher learning, institutional isomorphism, Latin America, Chile

Corresponding author

Rafael Sánchez
rafax@hotmail.com

* Rosario Rivero and María Eugenia Rojas gratefully acknowledge the financial support of Fondecyt Iniciación Nº11171096 and Support 2024 AFB240004 from the National Agency for Research and Development (ANID) respectively.

1. Introduction

Teachers are among the most influential school-based factors affecting student learning, yet the degree to which teacher preparation programs (TPPs) shape teacher effectiveness remains disputed. Early studies in the United States documented nontrivial differences across teacher preparation programs (TPPs) in graduates' value-added to student achievement (e.g., Boyd et al., 2009). However, a growing body of more recent evidence suggests that most variation in teacher effectiveness occurs within rather than between programs, and that program-level differences are generally modest in magnitude (Koedel et al., 2015; Mihaly et al., 2013; Bardelli, Ronfeldt, & Papay, 2023). At the same time, some studies indicate that preparation programs may differ in specific dimensions, such as early-career growth or contextual effectiveness, even when average differences are small. Against this backdrop, the Chilean case offers an opportunity to examine whether similarly compressed patterns of program differentiation emerge in a highly regulated middle-income system, where accountability and accreditation play a central role in shaping teacher education.

Chile offers a critical case. Over the past two decades, successive reforms have sought to elevate teacher quality through higher entry standards, mandatory accreditation, and accountability mechanisms (Brooks et al., 2022; Elacqua et al., 2018). Yet despite these efforts, concerns persist about whether TPPs adequately prepare teachers for the realities of diverse and inequitable classrooms. Persistent low performance on national teacher exams (INICIA) and the gap between policy ambition and practice raise a critical question: have regulatory reforms been associated with improvements in quality, or have they coincided with uniformity around mid-level performance?

This study examines whether Chilean TPPs differ in their contribution to teacher effectiveness and student achievement. Using national administrative data and value-added models,

we estimate the share of variance in student outcomes attributable to TPPs. Results show that TPPs explain only about 1% of the variance in student learning. Rather than evidence of uniformly strong preparation, this limited differentiation suggests that programs have converged toward a mid-level standard, reflecting compliance rather than excellence.

This study contributes to three strands of literature. First, it extends empirical evidence on teacher preparation program (TPP) effectiveness to a Latin American context where value-added analyses of preparation programs remain scarce. Second, it refines debates on program accountability by documenting limited differentiation across programs in a strongly regulated system. Third, it highlights the implications of institutional uniformity for teacher learning trajectories and educational equity. Together, these contributions advance cross-national discussions on how institutional design shapes the distribution of teacher effectiveness across preparation programs. To interpret the empirical patterns documented in this study, we draw on two complementary theoretical perspectives—institutional convergence and bounded instructional learning.

Empirically, the analysis relies on the only cohort for which Chilean administrative data allow a direct linkage between teachers, their preparation programs, and lagged student achievement. We follow students assessed in grade 8 in 2015 and again in grade 10 in 2017, linking their outcomes to novice mathematics teachers and the teacher preparation programs from which those teachers graduated. Consequently, the study does not exploit multiple cohorts or permit before–after comparisons around the 2016 reform. Instead, it provides a cross-sectional estimate of between- and within-program variation in teacher effectiveness at a specific point in time, which we then interpret considering recent regulatory developments in Chilean teacher education.

2. Materials and methods

2.1 Literature review on teacher preparation and effectiveness

A large body of research examines whether teacher preparation programs (TPPs) shape teacher effectiveness, often measured through teacher value-added to student achievement. In the United States, studies have documented modest but significant variation across programs (Boyd et al., 2009; Koedel et al., 2015; von Hippel et al., 2015), though other analyses attribute most variation to individual rather than institutional factors (Mihaly et al., 2013). Internationally, comparative studies suggest that preparation quality depends less on structural regulation than on opportunities for practice, mentoring, and feedback during training (Darling-Hammond et al., 2017; McDonald et al., 2013).

Despite this progress, empirical evidence from middle-income systems remains scarce. Most studies in Latin America rely on correlational designs linking teacher credentials or coursework to student outcomes (Ortúzar et al., 2009; Lara et al., 2010). Only recently have value-added approaches been applied to the region. For example, Santelices and Acuña (2019) and Canales and Maldonado (2018) used Chilean data to estimate teacher-level value-added, finding moderate effects of teacher training but limited between-program differentiation. These results mirror the broader international pattern: within-program heterogeneity dominates over between-program differences. However, no prior work has quantified how much of total variance in student learning can be attributed to Chilean TPPs themselves.

Unlike earlier Chilean studies that focused primarily on individual teacher characteristics or on single institutions, this analysis estimates the system-wide contribution of teacher preparation programs (TPPs) to student achievement using a multilevel value-added design. Prior research examined teacher-level effects within restricted samples of graduates but did not partition the total

variance in student learning attributable to TPPs as organizational units. In contrast, this study employs a three-level hierarchical model—students nested within teachers and teachers within programs—applied to a national dataset linking graduates from 30 TPPs to nearly 20,000 students. This design allows us to disentangle the institutional effect of programs from the individual effects of teachers. By incorporating both pretest scores and contextual covariates, the analysis improves identification relative to earlier correlational work that relied on cross-sectional associations between teacher credentials and student outcomes (Ortúzar et al., 2009; Lara et al., 2010). Finally, while prior studies assessed mean differences across teachers or compared graduates' test performance, none quantified the share of total variance in learning attributable to program-level factors versus within-program variation. Our study thus shifts the focus from isolated program evaluations to the systemic distribution of effectiveness across the teacher-education sector, providing the first national-level estimate of between- and within-program heterogeneity in Chile.

2.2. Conceptual framework

To interpret the empirical patterns documented in this study, we draw on four interrelated theoretical perspectives that inform contemporary research on teacher preparation and effectiveness. Given the scope and limits of the available data, these perspectives are not introduced as empirically tested mechanisms, but rather as theoretical lenses that help make sense of the observed structure of variation across and within teacher preparation programs (TPPs). In particular, they provide a framework for interpreting why program-level differentiation may be limited even in the presence of substantial within-program heterogeneity in teacher effectiveness.

The first perspective centers on teacher learning trajectories, which conceptualize teaching expertise as developing through iterative cycles of practice, feedback, and reflection (Ball & Forzani, 2009; Grossman et al., 2009). From this view, teacher preparation programs contribute to

effectiveness not only through formal coursework, but through the extent to which they embed sustained opportunities for clinically grounded learning. Differences in such opportunities, both across and within programs, may translate into variation in novice teachers' instructional effectiveness.

The second perspective draws on research on accountability, regulation, and institutional convergence. Organizational theories of institutional isomorphism suggest that, under strong regulatory and accreditation pressures, organizations operating within the same field tend to adopt similar structures, curricula, and practices (DiMaggio & Powell, 1983). In teacher education systems characterized by centralized standards and high-stakes accreditation, such pressures may reduce differentiation across preparation programs by aligning them around common requirements and compliance-oriented practices.

A third and related perspective emphasizes bounded instructional learning, which highlights how institutional constraints can shape and potentially limit the learning opportunities available to novice teachers. When preparation programs prioritize compliance with formal requirements over the enactment of practice-based learning, opportunities for developing robust instructional repertoires may be uneven or constrained. As a result, variation in teacher effectiveness may emerge primarily within programs, reflecting differences in practicum placements, mentoring relationships, and local school contexts rather than systematic program-level design features.

Finally, research on equity-oriented teacher preparation emphasizes the role of preparation programs in equipping teachers to work effectively in socially and economically diverse classrooms (Cochran-Smith & Villegas, 2015; Guillen & Zeichner, 2018). In stratified education systems, uniform program structures may mask important differences in how teachers are prepared to address inequality and diversity in practice. From this perspective, limited differentiation across programs

does not necessarily imply equitable preparation but may instead reflect shared constraints in addressing equity-focused instructional challenges.

Taken together, these perspectives provide an interpretive framework for situating the empirical findings of this study within broader debates on teacher learning, accountability, and institutional design, without claiming direct causal identification of specific mechanisms.

2.3. Contributions

Overall, this study contributes to international debates on teacher preparation effectiveness in three ways. First, it extends empirical evidence on teacher preparation program (TPP) effectiveness to a Latin American context where value-added analyses of preparation programs remain scarce. Second, it informs debates on program accountability by documenting limited differentiation across programs in a strongly regulated system, highlighting the potential for regulatory alignment to compress observable differences without necessarily improving overall quality. Third, it links teacher learning theory with system-level regulation by offering a theoretically grounded interpretation of how institutional design may shape the distribution—rather than the average level—of program effectiveness. Together, these contributions position Chile as a "critical case" (Flyvbjerg, 2006) for examining the limits of regulation-based approaches to improving teacher education quality.

2.4. Mechanisms and Theoretical Framework

The limited differentiation observed across teacher preparation programs (TPPs) in Chile is interpreted in this study through two complementary theoretical perspectives (institutional convergence and bounded instructional learning) operating within a strongly regulated policy

environment. Consistent with the conceptual framework outlined above, these perspectives are not presented as empirically tested mechanisms, but rather as interpretive lenses for understanding the observed patterns of low between-program variance and substantial within-program heterogeneity in teacher effectiveness.

From an institutional convergence perspective, organizational theories of institutional isomorphism suggest that, under strong regulatory and accreditation pressures, organizations operating within the same field tend to adopt similar structures, curricula, and practices (DiMaggio & Powell, 1983). In teacher education systems characterized by centralized standards, national entry requirements, and high-stakes accreditation, such pressures may reduce differentiation across preparation programs by aligning them around common requirements and compliance-oriented practices. From this perspective, accountability can standardize minimum conditions and inputs without necessarily expanding the distribution of program quality. At the same time, such regulatory convergence may serve as a safeguard against the risks associated with unregulated, low-quality teacher preparation, particularly in systems that previously lacked enforceable standards.[1]

A complementary perspective emphasizes bounded instructional learning, which highlights how institutional constraints can shape—and potentially limit—the learning opportunities available to novice teachers. Drawing on teacher learning as practice theory (Ball & Forzani, 2009; Grossman et al., 2009), teacher preparation is conceptualized as a process of developing instructional repertoires through rehearsal, feedback, and reflection. When accreditation and accountability emphasize formal inputs (such as course structure or contact hours) over enacted practice, opportunities for authentic clinical learning may be uneven or constrained. As a result, variation in

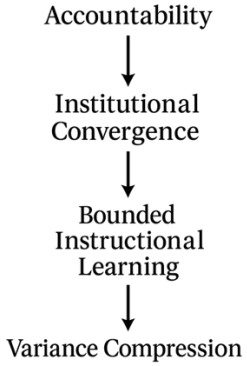
[1] For example, the National Academy of Education (2024) in the US highlight that standards and accountability can help raise the floor of program quality but must be balanced with efforts to promote innovation and equity. Similarly, OECD (2019) reviews emphasize that strong regulatory frameworks are essential to prevent low-quality teacher preparation, especially in systems with high levels of school and student diversity.

teacher effectiveness may emerge primarily within programs, reflecting differences in practicum placements, mentoring relationships, and local school contexts rather than systematic program-level design features.

Together, these perspectives provide a coherent interpretive framework for situating the Chilean case within broader theories of educational effectiveness. Rather than identifying causal effects of specific reforms, they help explain how policy design and institutional context may shape the distribution of teacher effectiveness observed in the empirical analysis. Figure 1 summarizes this interpretive model.

Figure 1

Conceptual model linking accountability reforms to teacher preparation program (TPP) effectiveness through institutional and instructional mechanisms.



This figure illustrates an interpretive model in which accountability and regulatory environments are associated with institutional convergence and bounded instructional learning, contributing to variance compression, that is, reduced differentiation across programs without substantive improvement in average performance.

2.5.Background

In the 80's, the implementation of market-based mechanisms to regulate the education system had a significant impact in the quality and social recognition of the teaching profession (Bellei, 2015). At the higher education level where teacher preparation programs operate, private providers were encouraged to create colleges, and vocational institutes became eligible for hosting teacher preparation programs. In this context, the emergence of a deregulated teacher preparation sector, where TPPs operated without standards of quality and with low entrance and exit requirements, contributed to the de-professionalization of the teaching workforce and the decrease in social esteem (Donoso, 2008). Most of the deregulation continued until recently. For example, up until 2015, only institutions receiving public funding, which represented 51% of all TPPs had to comply with mandatory accreditation (Santiago et al., 2017).

In response to quality concerns, during the last decade, Chilean reformers have designed and implemented several policies aimed at improving the training of teachers, including regulating entry requirements into teaching preparation programs, providing incentives to obtain graduate degrees, and subsidizing teacher preparation programs for institutions of higher education (Ávalos, 2010, 2014; Ávalos & Aylwin, 2007; Centro de Políticas Públicas, 2012; Cox, 2003; Cox et al., 2010, 2014; García-Huidobro, 2011; Mineduc, 2005; OECD, 2004; UNESCO, 2004; Ávalos and Valenzuela, 2016)[2]. In addition, Chilean policies have taken steps toward holding teacher preparation programs (TPPs) accountable for teacher quality.

[2] Teaching Vocation Scholarship (a national scholarship program to attract high-achieving students into the teaching profession), Teacher Initial Training Strengthening Program (FFID, a government initiative to improve pre-service

One of the most significant recent reforms is the Law 20,903 that creates the System for Teacher Professional Development (*Sistema de Desarrollo Profesional Docente*), enacted in 2016 and expected to be fully implemented by 2026. The main objective is for all teachers working in schools receiving public funding to become part of the System for Teacher Professional Development. Among other entitlements, with the passing of Law 20,903, become mandatory for teaching students to take a knowledge test at the beginning of their preparation program and at the end with the aim of assessing the impact of the programs (Santiago et al., 2017). Additionally, the Law established strict entry requirements for enrolling in TPPs: either obtain a minimum of 500 point in the national test for University entry, or be in the top 30% of the graduating high school class, or to have passed an accredited program for higher education access (Santiago et al., 2017). Complementary, Chile implemented national standards, listed in Table 1, that all TPPs must adopt in order to receive accreditation.

Table 1

Criteria and Standards for Teacher Education Programs in Chile.

Domain A. Teaching and learning preparation
Standard 1: pupils' learning and development
Standard 2: disciplinary, didactic, and school curriculum knowledge
Standard 3: lesson planning
Standard 4: assessment planning

teacher education), Mecesup projects (competitive funding for higher education improvement), the development of guiding standards for teacher education, the implementation of performance agreements between universities and the Ministry of Education, and the Teacher Career Law (Law 20.903, which established a national system for teacher professional development).

Domain B. Learning environment
Standard 5: organized and respectful environment
Standard 6: personal and social development
Domain C. Teaching for all
Standard 7: teaching strategies for deep learning
Standard 8: teaching strategies for thinking skills
Standard 9: assessment and feedback
Domain D. Professional responsibilities
Standard 10: professional ethics
Standard 11: professional development
Standard 12: commitment with the school community's improvement

Source: Brooks et al. (2022)

The accreditation process also required TPPs to cover four training areas: general (social and cultural factors; the education system; ethics and responsibilities); specialized (disciplinary knowledge; curricular content); professional (learning and teaching methods; tools for teaching) and practical (practice in schools) (Santiago et al., 2017). Finally, it is relevant to point out that even though there is important variability regarding the initial teacher preparation opportunities available, in Chile most teachers enroll in five-year programs similar to the ones pursued by other college graduates (Elacqua et al., 2018) and since 2014 teacher education is reserved exclusively to universities (Santiago et al., 2017).

This historical overview provides essential context for interpreting the regulatory environment in which the empirical analysis is situated, but the present study does not directly evaluate the causal effects of specific reforms or compare outcomes across regulatory regimes.

3. Results

3.1. Data Sources

This study uses Chilean students' achievement data from Simce[3]. Simce test results are crucially relevant for the school system in several ways. Simce receives intense publicity in the media stimulating discussion about educational quality, gender gaps, and social inequality. School administrators and teachers must devise mandatory improvement plans based on Simce test scores, and systematic underperforming schools risk closure. The Chilean Education Quality Agency (Agencia de la Calidad de la Educación) uses the test results as input for elaborating a school league table that informs resource allocation and public recognition. Parents can access historical school test performance data on the agency's webpage for accountability and school-choice information.

SIMCE assesses approximately 250,000 students per year, covering nearly all students in the grades selected for assessment. Standardized tests are administered to grades four, eight, and twelve according to a rotating calendar, such that not all grades and subjects are assessed every year. Across cycles, the assessments cover mathematics, language, sciences, and history. The recent test schedule produced only one cohort with lagged test scores available for our analysis. In this study, we use data from students first tested in grade eight (2015) and then in grade ten (2017). Although student achievement data are available for both mathematics and reading, we focus the analysis on mathematics because the data from teacher training programs are only available for mathematics; there are no data from teacher training programs in language or other specializations.

The Simce program employs criterion-referenced tests anchored in the Chilean mandatory national curriculum. The test scoring relies on an item response model approach with the person

[3] Simce is a standardized testing program managed by the Chilean Education Quality Agency.

ability estimates centered around 250 with a standard deviation of 50. Publicly available information reports a reliability/precision coefficient of 0.91 for the math test scores and 0.88 for the reading test scores (Simce, 2014; 2015). Our students' data includes five variables at the classroom-school level from Simce; 1) the average math and reading test scores from grade eight; 2) a socioeconomic status index (low, mid-low, mid, mid-high, high) built by the Education Quality Agency from student's background information (parents' income, parents' years of schooling, and an index of students' vulnerability); 3) type of school administration, public ownership and funding (public), private ownership with private and public funding (subsidized), and private ownership with private funding (private); 4) the school location (rural, urban). Table 2 provides descriptive statistics for all these five variables for each TPP.

Table 2

Student and School Variables Descriptive Statistics by TPP

TPP ID	Median Math Pretest	SD Math Pretest	Median Reading Pretest	SD Reading Pretest	Median Math Posttest	Math Posttet SD	SES Median	SES SD	Rural Schools %	Public School %	Subsidized Schools %	Private Schools %
1	249	39	220	43	265	49	2	1	0	9	82	9
2	291	32	254	64	246	65	3	1	0	20	80	0
3	255	65	257	56	248	87	2	1	0	40	50	10
4	242	52	246	39	227	55	2	1	3	47	47	6
5	316	62	309	58	344	81	5	1	0	0	17	83
6	233	50	235	47	232	75	3	1	9	36	64	0
7	245	45	224	49	231	54	2	1	8	30	68	3
8	244	44	214	51	249	54	2	1	0	14	84	2
9	283	58	278	37	270	65	3	2	0	14	43	43
10	255	43	262	42	204	48	3	1	0	40	60	0
11	279	43	266	43	291	62	3	1	0	26	65	9
12	280	46	257	48	282	65	2	1	8	13	82	5
13	275	28	268	32	279	50	3	1	0	44	44	11

14	247	45	238	47	245	50	2	1	0	43	57	0
15	250	39	245	38	251	54	2	1	8	49	49	3
16	263	41	253	39	277	54	2	1	6	38	62	0
17	274	16	228	24	290	79	3	1	0	100	0	0
18	221	17	229	26	208	50	1	1	0	100	0	0
19	215	52	224	16	255	64	1	1	0	57	43	0
20	286	38	258	36	295	59	3	1	0	39	48	12
21	252	47	253	45	238	59	2	1	2	40	58	2
22	267	41	239	46	234	50	1	1	17	46	54	0
23	274	40	283	53	299	41	3	1	0	40	60	0
24	250	50	237	45	255	59	2	1	8	38	56	7
25	267	42	241	51	288	58	3	1	0	27	65	8
26	247	49	196	44	172	64	1	1	0	20	80	0
27	275	47	271	45	308	49	4	1	0	20	80	0
28	268	57	249	56	241	61	2	1	13	41	59	0
29	279	54	251	42	263	71	3	1	9	36	64	0
30	253	43	249	44	236	54	2	1	4	32	68	0

This paper also uses teachers' data. We analyzed graduation records from TPPs across the country. Also, we collected information linking in-service teachers and the schools and classrooms they serve from public datasets available on the Chilean Ministry of Education webpage. This data enabled us to track novice math teachers in individual classrooms by the time of the Simce test administration. We identify 624 teachers from 30 TPPs linked to 26,853 students in 845 classrooms and 532 schools. Our 30 TPPs represent about 80% of all math TPPs in the country, and this group of novice teachers represents 85% of the total math teachers who graduated in the same period. Variation in the number of teachers per TPP is large. We traced between two teachers per program and 55 teachers per program, with a median of 31 teachers (SD=13.99). After dropping observations with incomplete students' test data (i.e., no unique student identification, no valid test score), the dataset contains information for 584 math teachers from 30 TPPs linked to 19,155 students in 845 classrooms and 532 schools. For analysis, we retained observations from classrooms with five or more students, reaching 576 math teachers from 30 TPP, linked to 19,131 students with valid test data in 835 classrooms and 524 schools.

Table 3

Students' Data Summary Statistics by TPP Information

TPP ID	Median Math Pretest	SD Math Pretest	Median Lan Pretest	SD Lan Pretest	Median Math Posttest	Math Posttest SD	SES Median	SES SD	Rural Schools %	Public School %	Subsidized Schools %	Private Schools %
1	249	39	220	43	265	49	2	1	0	9	82	9
2	291	32	254	64	246	65	3	1	0	20	80	0
3	255	65	257	56	248	87	2	1	0	40	50	10
4	242	52	246	39	227	55	2	1	3	47	47	6
5	316	62	309	58	344	81	5	1	0	0	17	83
6	233	50	235	47	232	75	3	1	9	36	64	0
7	245	45	224	49	231	54	2	1	8	30	68	3
8	244	44	214	51	249	54	2	1	0	14	84	2
9	283	58	278	37	270	65	3	2	0	14	43	43
10	255	43	262	42	204	48	3	1	0	40	60	0
11	279	43	266	43	291	62	3	1	0	26	65	9
12	280	46	257	48	282	65	2	1	8	13	82	5

TPP ID	Median Math Pretest	SD Math Pretest	Median Lan Pretest	SD Lan Pretest	Median Math Posttest	Math Posttest SD	SES Median	SES SD	Rural Schools %	Public School %	Subsidized Schools %	Private Schools %
13	275	28	268	32	279	50	3	1	0	44	44	11
14	247	45	238	47	245	50	2	1	0	43	57	0
15	250	39	245	38	251	54	2	1	8	49	49	3
16	263	41	253	39	277	54	2	1	6	38	62	0
17	274	16	228	24	290	79	3	1	0	100	0	0
18	221	17	229	26	208	50	1	1	0	100	0	0
19	215	52	224	16	255	64	1	1	0	57	43	0
20	286	38	258	36	295	59	3	1	0	39	48	12
21	252	47	253	45	238	59	2	1	2	40	58	2
22	267	41	239	46	234	50	1	1	17	46	54	0
23	274	40	283	53	299	41	3	1	0	40	60	0
24	250	50	237	45	255	59	2	1	8	38	56	7
25	267	42	241	51	288	58	3	1	0	27	65	8
26	247	49	196	44	172	64	1	1	0	20	80	0

TPP ID	Median Math Pretest	SD Math Pretest	Median Lan Pretest	SD Lan Pretest	Median Math Posttest	Math Posttest SD	SES Median	SES SD	Rural Schools %	Public School %	Subsidized Schools %	Private Schools %
27	275	47	271	45	308	49	4	1	0	20	80	0
28	268	57	249	56	241	61	2	1	13	41	59	0
29	279	54	251	42	263	71	3	1	9	36	64	0
30	253	43	249	44	236	54	2	1	4	32	68	0

Note: Cruch = older private institutions; CNA years = years of accreditation by the Chilean Quality Assurance Committee; PSU Min Score = college admission test score cutoff

3.2. Modeling Approach

Value-added modeling (VAM) is a quasi-experimental analytical strategy devised for capturing the contribution of programs or interventions on educational outcomes (Brick & Weisberg, 1976; Bryk & Weisberg, 1977; Bryk et al., 1980). VAM offers a more accurate estimation of treatment effect than the simple mean comparison (i.e., ANOVA) or adjusted mean comparison (i.e., ANCOVA) because mean and adjusted mean comparisons ignore the data generation process underlying students' achievement. VAM includes the knowledge about the effect of prior students' achievement on current students' achievement and uses the student as her/his own control. The value-added or treatment effect is the difference between the expected outcome had the program or intervention not occurred and the actual outcome. Thus, VAM requires at least one set of pretest scores. Pretest scores are a key component in our modeling approach. Pretests are also one of the best available statistical methods for controlling initial differences due to the non-random assignment of students to teachers (Guarino et al., 2014; Koedel et al., 2015; von Hippel, 2016). In some cases, pretest scores are the single most efficient source of statistical control. Student's background information such as the parent's educational level, income, or whether the student is a free-meal recipient may add little extra information (Wright et al., 1997). We estimated teacher and TPP value-added using a three-level random intercept model (student as level-1, teacher/classroom as level-2, and TPP as level-3) with lagged scores (pretest and posttest). The structural model for estimating the teacher´s contribution to his/her students' achievement using our lagged test data is the following:

$$Y_{ijk} = \gamma_{000} + \gamma_{100} Math_{ijk} + \gamma_{20k} Reading_{ijk} + \gamma_{010} C_{jk} + V_{00K} + U_{0jk} + R_{ijk} \qquad (1)$$

In equation 1, the math posttest score for student i with teacher j from TPP k (y_{ijk}) is a linear function of the grand posttest mean γ_{000}, the student's math and reading pretest scores, a set of classroom and school variables capturing influences beyond the control of the teacher (summarized as C_{jk}), and three error terms: V_{00K} capturing TPP-specific differences from the grand mean, U_{0jk} capturing classroom/teacher specific differences around the grand mean, and R_{ij} capturing student-level residuals not explained by the components in the model. The covariates (C_j) are the group test average in math and reading at pretest time, the students' socioeconomic level, type of school administration, and school location. We fitted several models introducing the covariates hierarchically. The model assumes that the variance components follow a normal distribution with a mean of zero.

We expect students from the same math teacher to show a more similar test performance than students from different math teachers due to the shared learning conditions, including teaching. Thus, the teacher-level random intercept U_{0jk} captures math test-scores heterogeneity explained by the teacher and not accounted for the covariates in the model. Likewise, we expect that test performance from students with teachers from the same TPP looks more alike than the test performance from students with teachers from different TPP. Thus, a TPP random intercept (V_{00K}) captures the heterogeneity in test scores accounted by the TPP and not accounted for the other variables in the model. We believe the TPP effect is the strongest during the first years of teaching and may decrease as the teacher gains professional experience. For this reason, our study only includes novice math teachers, and we excluded classrooms with more experienced teachers from our analysis. Also, our data only allows us to link novice teachers to his/her specific TPP and his/her students' test score and does not allow us to link a experienced teacher with his/her TPP nor his/her students' test scores.

Studies often debate between choosing a fixed effects (FE) model versus random effects (RE) model to capture TPP effects (von Hippel et al., 2016). We note that the likelihood of obtaining biased estimates depends upon the specific conditions of the study and that both FE and RE models need to comply with equally strong assumptions and the risk of incorrect model specification (Bell & Jones, 2015).

The choice of a random effect model in our study rests upon substantive and methodological reasons. First, a RE model allows the examination of the distribution of TPP effects (Bell & Jones, 2015) in direct connection with our research questions. Second, VAM includes random effects to account for the nested structure of educational data (i.e., students clustered within classrooms). Third, our data fails to meet the requirement of connecting all TPPs through at least one graduate teacher in a school with graduates from other programs (Mihaly et al., 2013). Literature on TPP effects recommends RE instead of a FE model with clustered standard errors when the number of clusters is small as in our case (von Hippel et al., 2016; 2018; Mihaly et al., 2013). Lastly, a random intercept model serves the purpose of examining heterogeneity across and within TPPs because programs and teachers are variance components directly embedded into the modeling approach.

The meaning of the teacher effects depends upon the set of covariates in the model. Covariates define the expected outcome. A model with only pretest scores produces estimates that contain the teacher effect confounded with peer, contextual and social effects. A model with no peer, contextual or social effects produces teacher effects that reflect the TPP ability to place teachers in favorable working contexts. Following literature on teacher VAM (Koedel et al., 2015; Chetty et al., 2014), we include prior achievement in math and language as well as average classroom prior achievement scores. The inclusion of students' peer and context effects in our

model results in estimates that allow meaningful comparisons across different settings (Willms & Raundenbush, 1989). Our list of covariates enables the interpretation of the value-added estimates as the net contribution of teachers (and TPP) because includes the effect of the school setting (e.g., principal's leadership, organizational culture, human and economic resources). We provide results with only pretest as covariate (without the term C_{jk}) and with the full set of covariates described above.

There are three steps in our analysis. First, we compare the relative fit across three models. A model 1 without random intercepts reflects the hypothesis of homogeneity across teachers and programs. Model 2 with teacher random intercept reflects the hypothesis of heterogeneity across teachers. Model 3 with teacher and TPP random intercepts reflects the hypothesis of heterogeneity across teachers and TPP. We use a likelihood ratio (LR) test to determine the best relative fitting model. A statistically significant LR test would provide empirical support to the unrestricted model over the simpler restricted model, favoring the hypothesis of heterogeneity across teachers and TPP. A non-statistically significant LR would provide empirical support to the restricted model in favor of the homogeneity across teachers and TPP. We replicated these three models twice: first with no covariates (only posttest scores), and then with covariates.

Then we examine the heterogeneity between and within TPP and report unexpectedly high and low value-added scores. The traditional method in textbooks and empirical research (e.g. Snijders & Bosker, 1999; Rabe-Hesketh & Skrondal, 2012) implies "predicting" the teacher and TPP random intercept using empirical bayes (EB) prediction. EB is also referred to as the best linear unbiased predictions (BLUP). We produce EB posterior means for U_{0jK} (teacher random intercept) and V_{00K} (TPP random intercept). EB estimation shrinkage construction provides a conservative view of the teacher and TPP value-added because posterior means are biased towards

the mean. Some authors argue that the EB shrinkage is problematic, especially in low value-added teachers (or TPPs) with a small number of observations (von-Hippel et al., 2015). However, in our study, we prefer estimates based on a small number of observations to have less influence on the overall results. Also, we prefer smaller prediction errors from EB over the unbiased but larger prediction errors using OLS and a teacher fixed effect (Snijders & Bosker, 1999, Rabe-Hesketh & Skrondal, 2012).

Following the prediction of value-added scores, a second step in the traditional contrast comprises building confidence intervals (CI). The CI allows identifying units with value-added predictions statistically above or below the average. We present graphical representations of between TPPs and within TPPs (across teachers) value-added distributions providing EB prediction point estimates with their CI using boxplots.

In summary, in answering our questions about Chilean math TPPs heterogeneity, we report estimates from two sets (with and without covariates) of three models (restricted, only teacher random intercept, teacher and TPP random intercept). For each model, we interpret FE and RE, intraclass correlation, model fit indexes and LR test. Then, we interpret value-added heterogeneity between TPPs and within TPPs using the EB prediction with their CI along with reliability of the estimates.

3.3. Results

3.3.1. How heterogenous are Math TPP?

Table 4 and 5 present FE (math test score grand mean) and RE estimates (teacher and TPP random intercepts), ICC, and fit information for our models testing the three hypotheses in the study: homogeneity across teachers and TPPs (model 1), heterogeneity across teachers (model 2),

and heterogeneity across teachers and TPPs (model 3). The first set of models (Table 4) ignores covariates while the second set of models (Table 5) includes covariates.

Table 4 reveals that the fixed effect part of the three models (model intercept) changes very little after including the random intercepts. The individual residual random effect estimates change dramatically from model 1 to model 2. The model 2 ICC reveals that 37% of the test scores variance corresponds to between teacher differences. The addition of the teacher random intercept results in a statistically significant improvement of the overall model fit (χ^2 (1) = 6457.70, p = 0.0000). The inclusion of a TPP random intercept in model 3 produces a less dramatic change. Specifically, the ICC suggests that 6% of the test scores variance corresponds to between TPPs differences. Despite this relatively small ICC, adding a TPP random intercept results in a statistically significant LR test (χ^2(1) = 57.70, p=0.000) and helps improve model fit.

Table 4

Only Posttest (Without Covariates) Model Estimates

	Model 1		Model 2		Model 3	
Parameter	Est.	Std Err.	Est.	Std Err.	Est.	Std Err.
FE						
Intercept	266.3	0.4	261.2	1.3	262.0	3.1
RE						
TPP					208.0	83.5
Teacher			1304.7	1304.7	1141.5	63.0
Residual	3499.8	35.8	2230.0	23.3	2230.0	23.3
ICC						
TPP					0.06	0.02
Teacher			0.37	0.01	0.38	0.02
FIT						
AIC	210413.5		203957.8		203902.1	
BIC	210429.2		203981.3		203933.5	

Note: model 1 vs model 2 LR χ^2 (1) = 6457.70, p = 0.0000; model 2 vs. model 3 LR χ^2 (1) = 57.70, p=0.000.

The same pattern emerges after incorporating covariates in our models. The fixed components (grand mean and regression coefficients) show small changes across the three models. In all the

models the size of the estimates for each fixed effect remains similar, with almost the same pattern of statistically significant coefficients. We found statistically significant coefficients for all but one covariate in model 1 (private school versus public school). The only difference between model 1 and models 2 and 3 relates to the lack of statistically significant difference between subsidized schools and public schools (Table 5).

The covariates show an impact on the RE estimates. Including covariates in the model reduces the residual variance in model 1 from 3499.8 (in Table 1) to 1426.49 (in Table 5). However, the proportion of variance accounted by teachers and TPPs across models replicates the pattern of findings about RE reported above. Differences between teachers account for the 14% of test scores variance, while the TPP random effect accounts for the 1% of test scores variance. The two LR tests resulted in statistically significant change in deviance, suggesting that the models with random intercepts offer a better relative fit than the models without random intercepts.

Table 5

Conditional Model Estimates (with Covariates)

	Model 1		Model 2		Model 3	
Parameter	Est.	Std Err.	Est.	Std Err.	Est.	Std Err.
FE						
Math pretest	0.72*	0.01	0.72*	0.01	0.72*	0.01
Language pretest	0.12*	0.01	0.12*	0.01	0.12*	0.01
Math average	0.22*	0.02	0.22*	0.05	0.20*	0.05
Language average	0.23*	0.02	0.21*	0.05	0.22*	0.05
Mid-low SES	10.03*	0.86	9.94*	1.68	9.70*	1.71
Mid SES	11.64*	1.04	12.41*	2.05	12.06*	2.09
Mid-high SES	16.23*	1.37	17.17*	2.76	17.08*	2.81
High SES	25.82*	3.13	24.46*	6.26	25.05*	6.21
Rural	10.60*	1.55	11.76*	2.87	11.14*	2.85
Subsidized school	2.11*	0.67	2.06	1.33	2.33	1.32
Private school	-3.69	3.19	-1.13	6.23	-1.09	6.16
Intercept	-84.58	4.15	-80.81	8.22	-77.80	8.23
RE						
TPP					12.03	6.03
Teacher			197.36	12.72	186.38	12.37

Residual	1426.49	14.59	1232.39	12.89	1232.42	12.89
ICC						
TPP					~~0.01~~	0.01
Teacher			~~0.14~~	0.01	~~0.14~~	0.01
FIT						
AIC	193265.3		191713.1		191701.6	
BIC	193367.5		191823.1		191819.5	

* Statistically significant, $p < 0.001$.

Note: model 1 vs model 2 LR χ^2 (1) = 1554.23, p = 0.0000; model 2 vs. model 3 LR χ^2 (1) = 13.45, p=0.0002.

Although the variance in student outcomes attributable to teachers and TPP are modest (14% and 1% respectively), this result has important theoretical implications. From a teacher learning perspective, such limited differentiation suggests that most programs are structuring teacher preparation along similar trajectories, offering few distinctive opportunities for extended practice, feedback, or reflection (Ball & Forzani, 2009; Grossman et al., 2009). In terms of equity-oriented preparation, program homogeneity may also imply that teachers are not systematically prepared to address the deep inequalities and cultural diversity of Chilean classrooms (Cochran-Smith & Villegas, 2015; Guillen & Zeichner, 2018). Finally, with regard to accountability, the small variance questions whether program-level evaluation systems based primarily on value-added scores can effectively capture meaningful differences in quality across TPPs (Koedel et al., 2015).

Our results suggest that the data conforms with model 3 supporting the hypothesis of heterogeneity across TPP. The heterogeneity across TPPs remains even after including covariates accounting by initial differences among students including of peer, social and contextual effects. The size of these differences across TPPs is rather small. Figure 2 portrays how much heterogeneity is due to programs versus teachers and students. The figures show that the major source of test scores variability is the students followed by teachers and then TPP.

3.3.2. How is the heterogeneity between and within TPP?

The following boxplot presents the value-added scores and 95% posterior confidence interval for each of the 30 math TPPs in our study. The graph reveals that most confidence intervals overlap the mean value-added, suggesting that the overall effectiveness of these programs are indistinguishable from one another. There are four programs with confidence intervals not overlapping the mean value-added, two at the bottom (IDs 4 and 21) and two at the top (IDs 14 and 15) of the value-added distribution.

Figure 2

Boxplot of Empirical Bayes Prediction for TPP And Teacher Random Intercepts and Individual Residuals

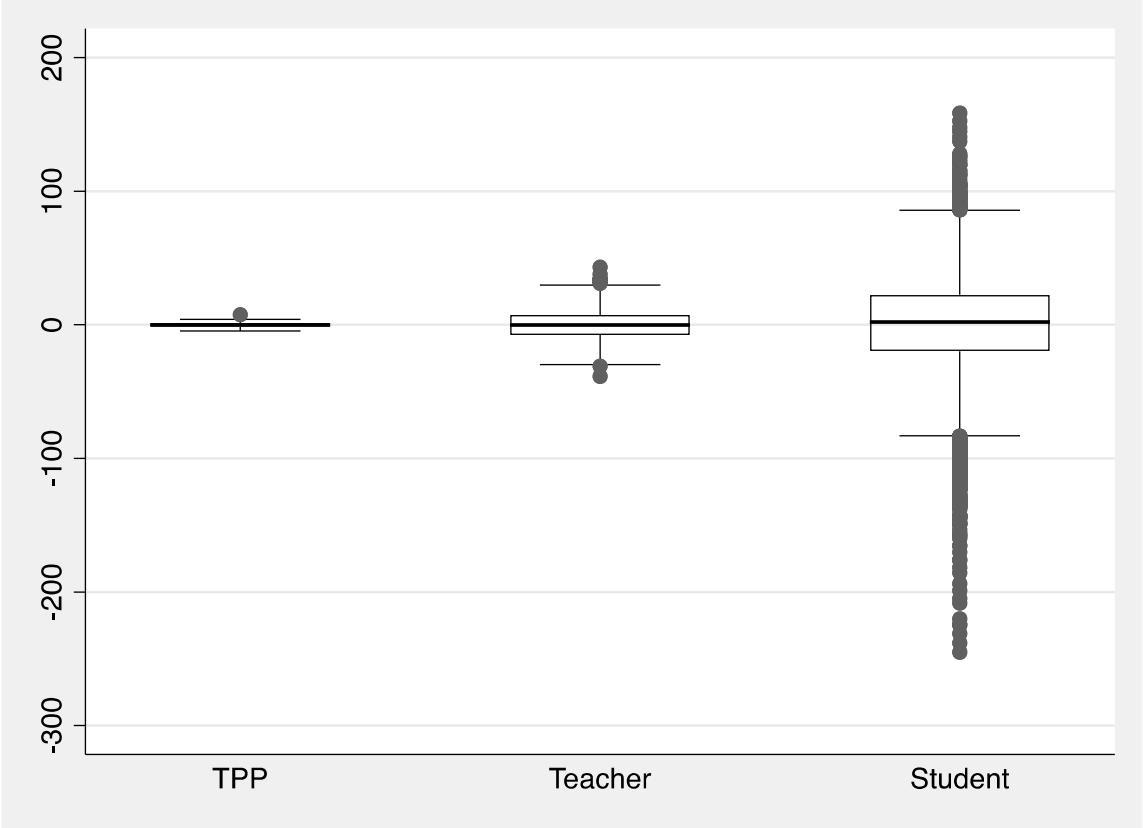


Table 6 presents teacher value-added summary statistics for all 30 TPP. From the results, we identified four types of programs based on the reliability and size of the teacher value-added scores: two programs with homogeneous below average teacher value-added (IDs 4 and 21), eight programs with homogeneous average teacher value-added (IDs 5, 7, 13, 19, 23, 26, 29, 30), fourteen programs with heterogeneous average teacher value-added (IDs 1, 2, 3, 6, 8, 11, 12, 16, 20, 22, 24, 25, 27, 28), and two programs with heterogeneous above average teacher value-added (IDs 14 and 15). The programs IDs 9, 10, 17 and 18 feature homogeneous teacher value added but estimates are based on less than five teachers.

The distribution of teachers' value-added in the two TPPs with overall below-average value-added (top half in figure 3) is homogeneous (with a reliability below 0.7) and comprises 43%-50% of teachers with a below-average value-added, 14%-25% with an average value-added, and 36%-32% with above average value-added, per program respectively. The two programs with an overall above-average value-added (bottom half in figure 3) show a fairly similar heterogeneous distribution of teacher value-added (reliability above 0.7). Almost a third (31%-32%) of the teachers show a below average value-added, another third of the teachers (31-32%) show an average value-added, and 35%-38% of the teachers show an above-average value-added, per program respectively.

Table 6

Summary Statistics of Teacher Value-added Scores for 30 Math TPP

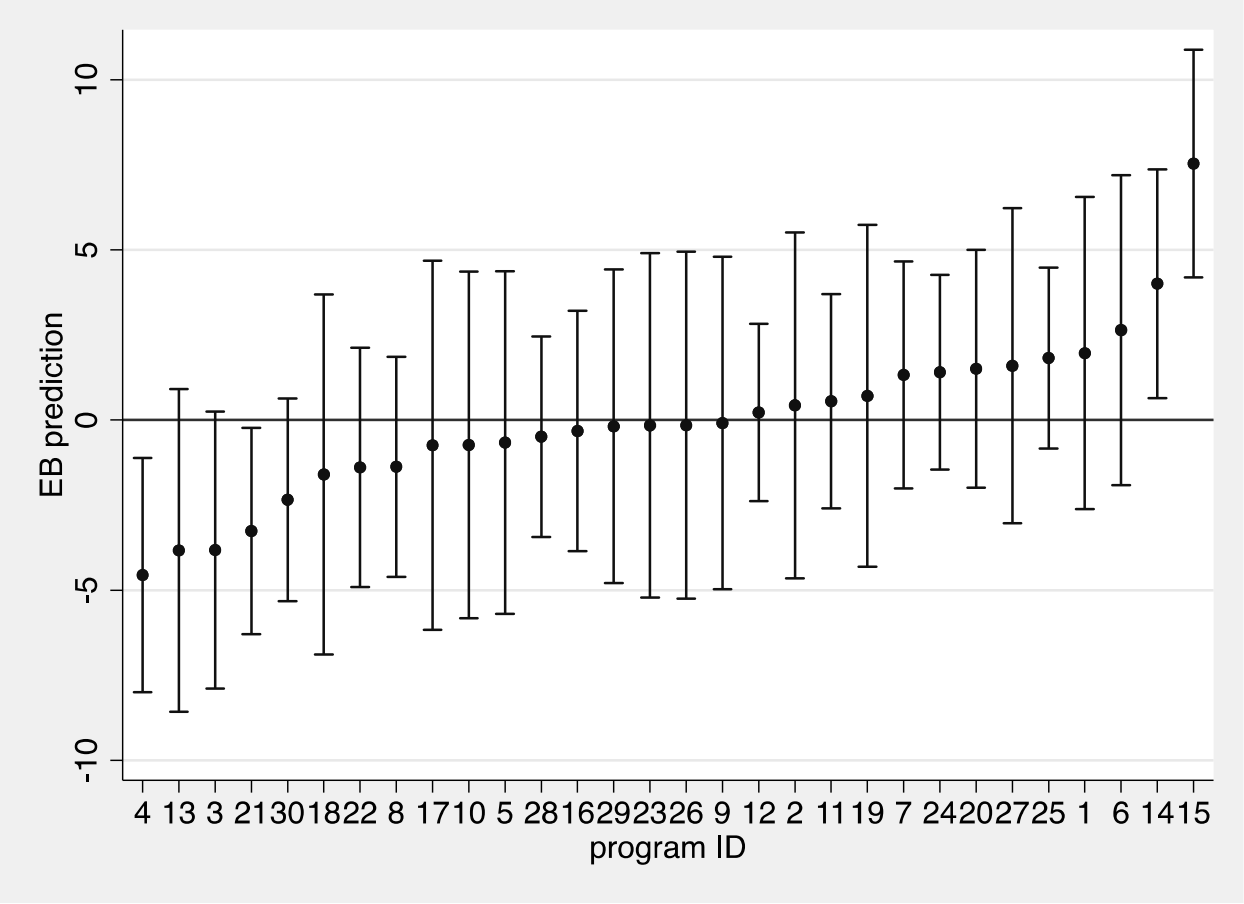
TPP ID	N Teachers	TPP VA Ranking	TPP VA Min	TPP VA Max	TPP VA Mean	TPP VA SD	Teachers Below VA %	Teachers Average VA %	Teachers Above VA %	TPP VA Reliability	TPP Type
1	11	27	-19.27	20.17	1.96	12.89	36	18	45	0.74	Heterogeneous - $\overline{VA}$
2	5	19	-10.75	16.01	0.43	13.31	40	20	40	0.75	Heterogeneous - $\overline{VA}$
3	20	3	-38.61	16.30	-3.82	13.68	45	20	35	0.76	Heterogeneous - $\overline{VA}$
4	36	1	-24.19	21.09	-4.56*	10.74	50	14	36	0.61	Heterogeneous - below $\overline{VA}$
5	6	11	-13.82	9.82	-0.66	8.84	33	33	33	0.29	Homogeneous - $\overline{VA}$
6	11	28	-14.51	33.27	2.64	15.92	45	18	36	0.85	Heterogeneous - $\overline{VA}$
7	40	22	-22.60	28.23	1.33	11.26	38	23	40	0.66	Homogeneous - $\overline{VA}$
8	43	8	-26.93	34.29	-1.37	12.89	35	30	35	0.76	Heterogeneous - $\overline{VA}$
9	7	17	-10.41	6.46	-0.09	6.79	29	57	14	0.00	Homogeneous[1] $\overline{VA}$
10	5	10	-12.41	5.62	-0.73	6.57	20	60	20	0.00	Homogeneous[1] $\overline{VA}$
11	46	20	-26.02	37.81	0.55	13.99	43	15	41	0.78	Heterogeneous - $\overline{VA}$
12	76	18	-29.81	20.03	0.22	9.60	42	21	37	0.53	Heterogeneous - $\overline{VA}$
13	9	2	-17.40	5.43	-3.83	8.28	56	33	11	0.36	Homogeneous - $\overline{VA}$

14	37	29	-24.09	33.82	4.01*	12.53	32	32	35	0.74	Heterogeneous - above $\overline{VA}$
15	39	30	-20.32	43.13	7.54*	12.89	31	31	38	0.75	Heterogeneous - above $\overline{VA}$
16	34	13	-30.98	29.77	-0.32	14.13	38	18	44	0.77	Heterogeneous - $\overline{VA}$
17	2	9	-7.28	-4.22	-0.74	2.17	50	50	0	0.00	Homogeneous[1] $\overline{VA}$
18	3	6	-14.53	-3.07	-1.60	5.80	67	33	0	0.00	Homogeneous[1] - $\overline{VA}$
19	7	21	-9.93	22.24	0.71	12.53	43	14	43	0.57	Homogeneous - $\overline{VA}$
20	33	24	-25.18	31.65	1.51	14.32	42	6	52	0.78	Heterogeneous - $\overline{VA}$
21	53	4	-21.35	20.41	-3.26*	10.01	43	25	32	0.54	Heterogeneous - below $\overline{VA}$
22	35	7	-18.78	23.13	-1.39	9.75	40	29	31	0.42	Heterogeneous - $\overline{VA}$
23	5	15	-6.38	5.53	-0.16	5.29	20	60	20	0.00	Homogeneous - $\overline{VA}$
24	61	23	-18.75	18.34	1.40	9.88	41	15	44	0.56	Heterogeneous - $\overline{VA}$
25	74	26	-25.95	22.87	1.82	11.34	35	19	46	0.66	Heterogeneous - $\overline{VA}$
26	5	16	-15.63	15.32	-0.15	11.11	20	60	20	0.48	Homogeneous - $\overline{VA}$
27	10	25	-16.88	34.02	1.59	15.75	50	10	40	0.83	Heterogeneous - $\overline{VA}$
28	54	12	-29.35	34.17	-0.49	14.42	46	19	35	0.79	Heterogeneous - $\overline{VA}$
29	11	14	-12.32	16.17	-0.19	8.33	27	55	18	0.30	Homogeneous - $\overline{VA}$
30	57	5	-25.74	22.83	-2.34	11.27	39	23	39	0.58	Homogeneous - $\overline{VA}$

* Posterior confidence interval does not overlap the mean value-added score; [1]Low number of teachers; Homogeneous VA from testing the null hypothesis that the heterogeneity variance equals 0 with the Cochran's Q statistics according to von Hippel & Bellows (2018.)

Figure 3

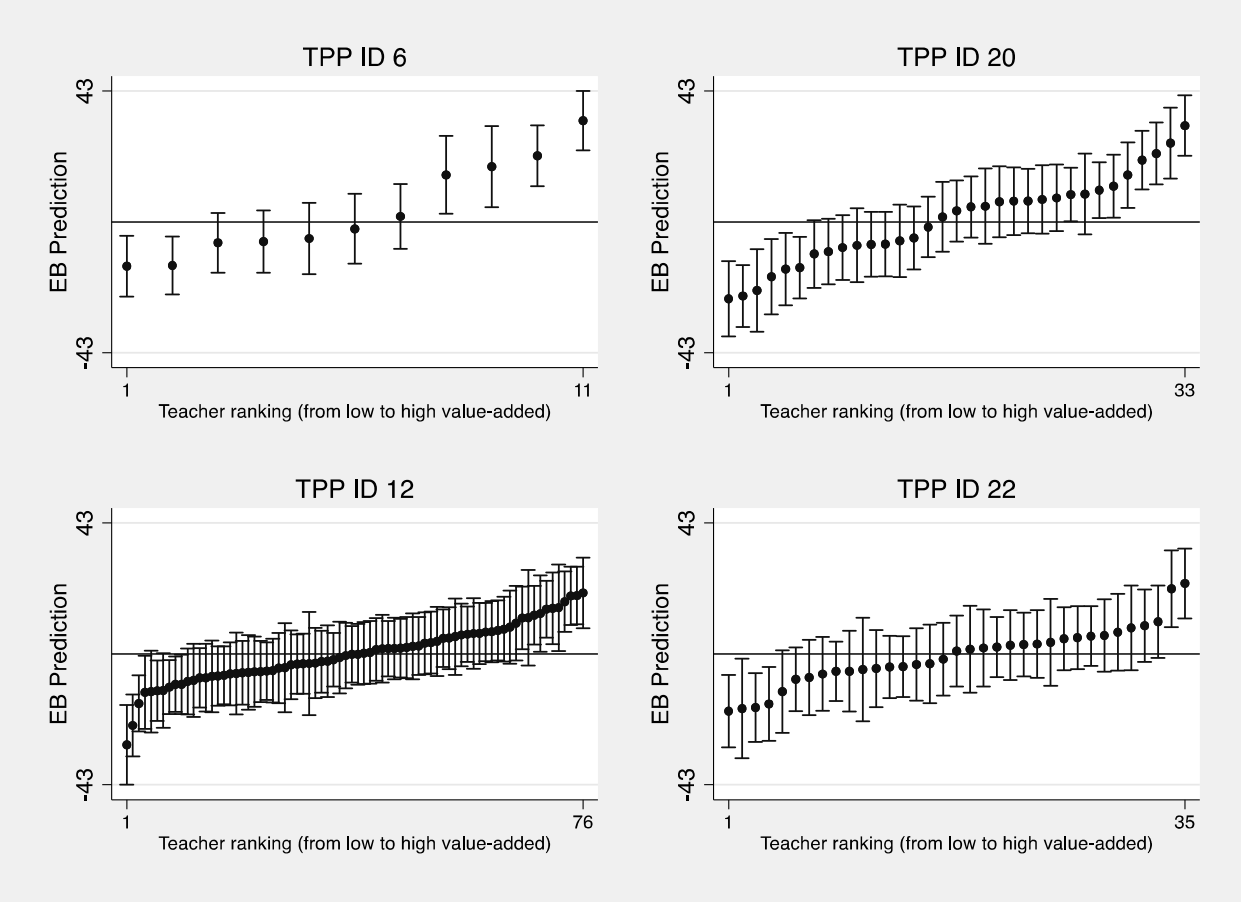
Value-Added Scores for 30 TPPs with 95% Posterior Confidence Intervals



The boxplots in figure 4 depict four types of programs with an average teacher value-added. Programs IDs 6 and 20 show heterogeneous teacher value-added distributions (reliabilities 0.85 and .78, respectively) with a small (11) and large (33) number of teachers. Teachers concentrate either at the top or bottom of the distribution, with only 18% and 6% of the teachers exhibiting an average value-added, respectively. On the contrary, programs IDs 12 and 22 show homogeneous distributions (reliabilities 0.53 and .42) both with a large number of teachers (76 and 35 respectively). The proportion of teachers with an average value-added in these programs is 21% and 29%, respectively.

Figure 4

Boxplot of Teacher Value-Added for Programs with Heterogeneous (Top) and Homogeneous (Bottom) Average Value-Added



4. Discussion

Our findings reveal limited differentiation across Chilean teacher preparation programs (TPPs): programs account for only about 1% of the variance in student achievement. This apparent homogeneity should not be mistaken for uniformly high quality. Rather, it reflects a convergence around mid-level performance, suggesting that regulatory alignment appears to have standardized inputs and structures, a pattern that may limit the expansion of learning opportunities for future

teachers. To interpret these results, we examine how two potential mechanisms identified in our theoretical framework—institutional convergence and bounded instructional learning—manifest in the empirical patterns observed

4.1. Potential mechanism 1: Institutional convergence and variance compression

The limited between-program variance observed, in this study, approximately 1% of total variance in student achievement, is consistent with processes of institutional convergence across teacher preparation programs (TPPs). In the Chilean context, two decades of regulatory reforms have required programs to align curricula, admission standards, and assessment procedures to centralized accreditation frameworks. Drawing on institutional isomorphism (DiMaggio & Powell, 1983), this pattern suggests that external accountability has standardized program structures and content, thereby reducing differentiation in measurable outcomes.

Rather than reflecting uniformly high quality, this convergence likely represents a compliance-driven form of homogenization. Programs optimize for accreditation success and regulatory approval rather than for pedagogical innovation or context-specific practice. The resulting variance compression explains why accountability can reduce the dispersion of effectiveness estimates without raising overall quality. The small interclass correlation coefficient (ICC = 0.01) thus functions as a quantitative indicator of isomorphic pressures in a tightly regulated system. This pattern, consistent with convergence underscores how regulation can homogenize institutional structures, a process further illuminated when we examine the variability that persists within programs.

4.2. Potential mechanism 2: Bounded instructional learning and within-program heterogeneity

Although between-program variation is small, the model reveals substantial within-program heterogeneity in teacher effects, consistent with the mechanism of bounded instructional learning—where institutional and curricular constraints limit opportunities for sustained practice-based learning during preparation.

Drawing on the teacher learning as practice framework (Ball & Forzani, 2009; Grossman et al., 2009), we interpret this internal variability as evidence that, despite formal standardization, pre-service experiences remain uneven across practicum placements, mentors, and feedback structures. Regulation has ensured procedural alignment but not the consistent enactment of learning opportunities that develop teaching expertise. The bounded nature of instructional learning—shaped by accreditation requirements, limited supervision, and scarce school–university partnerships—therefore explains the persistence of large within-program differences even under a uniform policy regime. While institutional convergence and bounded instructional learning are inferred from variance structures rather than observed directly, future mixed-methods research could trace how accreditation and mentoring practices mediate these quantitative patterns.

4.3. Implications for teacher preparation programs

While these implications are not directly tested in the present analysis, they follow from the observed variance patterns and from established evidence in the teacher learning literature. These findings suggest that future policy discussions may consider strengthening mentoring and clinical practice in teacher preparation programs, particularly in systems characterized by strong regulation.

4.3.1. Institutional mechanisms: Mentoring and induction

Weak links between universities and schools limit the transfer of learning from preparation to practice. Only a small fraction of novice teachers in Chile receive structured mentoring (Akiba

et al., 2022), which may hinder the consolidation of professional expertise. Future policy discussions may consider strengthening partnerships between universities and schools—through jointly trained mentors, shared evaluation frameworks, and continuous supervision—could extend learning across the pre-service and induction stages.

International examples such as Finland's research-based mentorship and Singapore's structured induction illustrate how coordinated systems can reinforce professional growth. Embedding such practices locally would bridge the preparation–practice gap and give regulatory frameworks substantive meaning.

4.3.2. System-level mechanisms: Accountability and convergence

Results highlight both the power and limits of accountability-based reform. Standardized accreditation and entry requirements have reduced program-level variance but may also have encouraged convergence toward mid-level performance. When compliance metrics emphasize formal requirements over instructional quality, programs optimize for minimum standards rather than innovation.

This dynamic exemplifies the broader paradox of educational regulation: it promotes transparency and comparability while discouraging differentiation. Reframing accountability to emphasize continuous improvement—through peer review, formative assessment of teaching practice, and feedback loops—could transform oversight into a driver of professional learning rather than standardization.

4.4. Implications and broader contribution

Methodologically, our results demonstrate that value-added models can illuminate not only program differences but also systemic convergence, revealing where policy reforms have compressed rather than expanded variation in quality. Substantively, the findings caution against

interpreting homogeneity as success: when all programs look alike, accountability may have achieved alignment but not advancement. The Chilean case thus contributes to a growing international literature questioning whether strong regulation necessarily yields stronger teachers (von Hippel et al., 2016; Snyder & Lit, 2023). More broadly, the study underscores that improving teacher effectiveness requires more than institutional compliance. Sustained progress depends on transforming TPPs into learning-centered organizations that integrate extended practice, mentorship, and equity-driven pedagogy. For countries in Latin America, this implies that convergence should be upward—toward excellence—rather than lateral, toward sameness.

4.5. Limitations and Future Research

Despite its contributions, this study has some limitations that warrant attention. First, while the value-added approach strengthens causal inference, residual selection and measurement error may still attenuate estimates of program-level effects. Second, the analysis captures one cohort and one policy period; future research could examine whether institutional convergence and bounded instructional learning persist or evolve under subsequent reforms. Third, we only focus on mathematics because the data from teacher training programs is only available for mathematics; there are no data from teacher training programs in language or other specializations. Fourth, the linkage between teachers, their teacher preparation programs, and their students' data is only available for novice teachers. Fifth, the mechanisms identified here are inferred from variance patterns rather than observed directly in program practices. Qualitative or mixed-methods studies could further unpack how regulatory pressures translate into learning opportunities within teacher preparation programs. Addressing these questions would deepen understanding of how accountability systems shape the distribution and dynamics of teacher effectiveness.

Conclusion

Our findings nuance the dominant narrative in international research that teacher preparation programs differ widely in their effectiveness. In contrast to early evidence from the United States, where substantial program-level differences have been documented (Boyd et al., 2009; Koedel et al., 2015), Chilean TPPs display remarkable homogeneity despite extensive reforms and strong regulatory frameworks. Yet this convergence does not reflect a collective movement toward excellence; rather, it signals alignment around a common, mid-level standard of preparation. This paradox—programs becoming more similar without necessarily becoming stronger— suggests questions for future research for the field of teacher education. It suggests that accountability and accreditation, if not paired with deliberate efforts to expand practice opportunities, strengthen equity-oriented curricula, and institutionalize mentoring, may generate uniformity without quality.

In sum, this study provides new evidence to a longstanding debate on the role of teacher preparation programs (TPPs) in shaping student outcomes, while offering international insights into how pre-service preparation influences teacher learning in highly regulated systems. Although program-level variance in teacher effectiveness is modest, the implications are nonetheless significant. Pre-service education remains a central component of teacher development, particularly in contexts such as Chile where opportunities for sustained, high-quality in-service professional learning are limited. Addressing this challenge requires more than regulatory compliance. It calls for reimagining TPPs as institutions that deliberately and systematically support teacher learning through extended clinical practice, equity-oriented curricula, structured mentoring, and continuous feedback. Grounding reform efforts in these principles would strengthen teacher education systems across Latin America and better prepare educators for classrooms increasingly shaped by inequality, diversity, and technological change. More broadly,

this study contributes to international debates by illustrating that homogeneity is not synonymous with effectiveness, and by underscoring the need for policies that promote convergence upward—toward excellence—rather than laterally, toward sameness. Accordingly, the study is designed to describe the distribution of teacher effectiveness across programs at a single point in time, rather than to estimate changes induced by specific reforms.

Acknowledgments

All listed authors made substantial contributions to the development of this manuscript. We used AI-assisted tools for language editing.

Statements and Declarations.

Ethical considerations. Not obtained because the data used in this study are secondary in nature and were provided in anonymized form by the Chilean Education Quality Agency (Agencia de la Calidad de la Educación).

Consent to participate. Not obtained because the data used in this study are secondary in nature and were provided in anonymized form by the Chilean Education Quality Agency (Agencia de la Calidad de la Educación).

Conflict of interest. None. The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding. Rosario Rivero and María Eugenia Rojas gratefully acknowledge the financial support of Fondecyt Iniciación Nº11171096 and Support 2024 AFB240004 from the National Agency for Research and Development (ANID) respectively.

Data availability. The data used in this study cannot be shared as it comes from administrative records provided by Chilean Education Quality Agency (Agencia de la Calidad de la Educación).

References

Akiba, M., Murillo, F. J., & Rojas, M. E. (2022). *Mentoring and induction support for beginning teachers in Latin America: Evidence from TALIS 2018. Teaching and Teacher Education, 112*, 103640.

Ávalos, B., & Aylwin, P. (2007). La formación inicial docente en Chile: 25 años de debates y reformas. *Revista Pensamiento Educativo, 41*(1), 17–40.

Ávalos, B. (2010). La formación inicial docente en Chile: Tensiones entre políticas de apoyo y control. *Revista Española de Educación Comparada, 16*, 235–263.

Ávalos, B. (2014). La formación inicial docente en Chile: avances, tensiones y desafíos. *Estudios Pedagógicos, 40*(Especial), 11–28.

Ávalos, B., & Valenzuela, J. P. (2016). Education for all and attrition/retention of new teachers in Chile. *International Journal of Educational Development, 49*, 279–290.

Ball, D. L., & Forzani, F. M. (2009). The work of teaching and the challenge for teacher education. *Journal of Teacher Education, 60*(5), 497–511.

Bardelli, E., Ronfeldt, M., & Papay, J. P. (2023). Teacher preparation programs and graduates' growth in instructional effectiveness. *American Educational Research Journal, 60*(1), 183–216.

Bell, A., & Jones, K. (2015). Explaining fixed effects: Random effects modeling of time-series cross-sectional and panel data. *Political Science Research and Methods, 3*(1), 133–153.

Bellei, C. (2015). *El gran experimento: Mercado y privatización de la educación chilena*. Santiago de Chile: LOM Ediciones.

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

Brick, J. M., & Weisberg, H. I. (1976). Problems in estimating the effect of treatment when subjects are nested within institutions. *Journal of Educational Statistics, 1*(2), 97–107.

Brooks, C., Bustos, T., Cox, C., Meckes, L., & Pino, M. (2022). *Standards for initial teacher education in Chile: Policy design and implementation. Journal of Education Policy, 37*(6), 781–802.

Bryk, A. S., & Weisberg, H. I. (1977). Value-added analysis for estimating school effects. *Sociology of Education, 50*(4), 241–258.

Bryk, A. S., Strenio, J., & Weisberg, H. I. (1980). Value-added analysis of educational institutions: External validity issues. *Sociological Methodology, 11*, 325–359.

Canales, A., & Maldonado, L. (2018). Teacher quality and student achievement in Chile: Assessing the role of value-added models. *International Journal of Educational Development, 60*, 33–50.

Centro de Políticas Públicas UC. (2012). *Formación inicial docente: estado del arte y desafíos.* Pontificia Universidad Católica de Chile.

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

Cochran-Smith, M., & Villegas, A. M. (2015). Framing teacher preparation research: An overview of the field, part I. Journal of Teacher Education, 66(1), 7–20.

Cox, C. (2003). *Políticas educacionales en el cambio de siglo: La reforma del sistema escolar de Chile.* Editorial Universitaria.

Cox, C., Meckes, L., & Bascopé, M. (2010). Calidad de la formación inicial docente en Chile: Base de un sistema de aseguramiento de la calidad. *Revista Pensamiento Educativo, 46*(1), 17–50.

Cox, C., Meckes, L., & Bascopé, M. (2014). La formación inicial docente en Chile: 1990–2010. *Revista Pensamiento Educativo, 51*(1), 17–40.

Darling-Hammond, L., Burns, D., Campbell, C., Goodwin, A. L., Hammerness, K., Low, E. L., McIntyre, A., Sato, M., & Zeichner, K. (2017). *Empowered educators: How high-performing systems shape teaching quality around the world.* Jossey-Bass.

DiMaggio, P. J., & Powell, W. W. (1983). The iron cage revisited: Institutional isomorphism and collective rationality in organizational fields. *American Sociological Review, 48*(2), 147–160.

Donoso, S. (2008). Políticas de formación inicial docente en Chile: 1990–2007. *Revista Pensamiento Educativo, 42*(2), 207–227.

Elacqua, G., Hincapié, D., Vegas, E., Alfonso, M., & Montalva, V. (2018). *Profesión: Profesor en América Latina. ¿Por qué se perdió el prestigio docente y cómo recuperarlo?* Inter-American Development Bank.

Flyvbjerg, B. (2006). Five misunderstandings about case-study research. *Qualitative Inquiry, 12*(2), 219–245.

García-Huidobro, J. E. (2011). Formación inicial de profesores en Chile: Estado actual y desafíos. *Revista Docencia, 45*, 5–16.

Grossman, P., Compton, C., Igra, D., Ronfeldt, M., Shahan, E., & Williamson, P. (2009). Teaching practice: A cross-professional perspective. *Teachers College Record, 111*(9), 2055–2100.

Guarino, C. M., Reckase, M. D., & Wooldridge, J. M. (2014). Can value-added measures of teacher performance be trusted? *Education Finance and Policy, 10*(1), 117–156.

Guillen, L., & Zeichner, K. (2018). A university-community partnership in teacher education from the perspectives of community-based teacher educators. *Journal of Teacher Education, 69*(2), 140–153.

Koedel, C., Parsons, E., Podgursky, M., & Ehlert, M. (2015). Teacher preparation programs and teacher quality: Are there real differences across programs? *Education Finance and Policy, 10*(4), 508–534.

Koedel, C., Mihaly, K., & Rockoff, J. E. (2015). Value-added modeling: A review. *Economics of Education Review, 47*, 180–195.

Lara, B., Mizala, A., & Repetto, A. (2010**).** The effectiveness of private voucher schools in Chile: Evidence from standardized test results. *Educational Evaluation and Policy Analysis, 32*(3), 219–243.

McDonald, M., Kazemi, E., & Kavanagh, S. S. (2013). Core practices and pedagogies of teacher education: A call for a common language and collective activity. *Journal of Teacher Education, 64*(5), 378–386.

Mihaly, K., McCaffrey, D. F., Sass, T. R., & Lockwood, J. R. (2013). Where you come from or where you go? Distinguishing between school quality and the effectiveness of teacher preparation program graduates. *Education Finance and Policy, 8*(4), 459–493.

Ministerio de Educación de Chile (Mineduc). (2005). *Política de formación inicial docente.* Santiago: Gobierno de Chile.

National Academy of Education. (2024). *Evaluating and Improving Teacher Preparation Programs* (K. M. Zeichner, L. Darling-Hammond, A. I. Berman, D. Dong, & G. Sykes, Eds.). National Academy of Education.

Organization for Economic Co-operation and Development (OECD). (2004). *Reviews of National Policies for Education: Chile 2004, Reviews of National Policies for Education, OECD Publishing, Paris.*

Organization for Economic Co-operation and Development (OECD). (2019). *A Flying Start: Improving Initial Teacher Preparation Systems*. OECD Publishing, Paris.

Ortúzar, M., Flores, C., Milesi, C., & Cox, C. (2009). Efectividad escolar y rol de los docentes en Chile: ¿Qué nos dice la evidencia? *Revista Pensamiento Educativo, 45*(1), 63–82.

Rabe-Hesketh, S., & Skrondal, A. (2012). *Multilevel and longitudinal modeling using Stata* (3rd ed.). College Station, TX: Stata Press.

Santelices, M. V., & Acuña, F. (2019). Is teacher education worth it? Value-added estimates for a Chilean teacher education program. *Teaching and Teacher Education, 78*, 106–121.

Santiago, P., Fiszbein, A., García Jaramillo, S., & Radinger, T. (2017). *OECD Reviews of School Resources: Chile 2017.* OECD Publishing.

Snijders, T. A. B., & Bosker, R. J. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling.* London: Sage.

Snyder, J., & Lit, I. (2023). Accountability and teacher preparation: Shifting the conversation toward learning opportunities. *Journal of Teacher Education, 74*(2), 211–224.

UNESCO. (2004). *Reformar la formación docente: desafíos y perspectivas en Chile.* Santiago: UNESCO.

von Hippel, P. T. (2016). Measures of student learning and teacher productivity. In H. Ladd & M. Goertz (Eds.), *Handbook of research in education finance and policy* (2nd ed., pp. 328–346). Routledge.

von Hippel, P. T., Lahiri, S., & Fitzpatrick, M. (2015). How much does teacher quality vary across teacher preparation programs? *Educational Evaluation and Policy Analysis, 37*(4), 792–810.

von Hippel, P. T., Bellows, L., Hargrove, D. T., & Nicholson-Crotty, S. (2016**).** Teacher quality and student learning in the United States: Evaluating the evidence. In M. Akiba & G. LeTendre (Eds.), *International Handbook of Teacher Quality and Policy* (pp. 57–74). Routledge.

von Hippel, P. T., Jellison, J., & Black, R. (2018). When does teacher value-added stabilize? Comparing statistical models with empirical data. *Economics of Education Review, 64*, 1–13.

Willms, J. D., & Raudenbush, S. W. (1989). A longitudinal hierarchical linear model for estimating school effects and their stability. *Journal of Educational Measurement, 26*(3), 209–232.

Wright, S. P., Horn, S. P., & 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.