Teachers' beliefs about standardised testing and test-based accountability: comparing the perceptions and experiences of teachers in Chile and Norway

(forthcoming in *European Journal of Education*)

Marjolein K. Camphuijsen[a]* and Lluís Parcerisa[b]

[a]*Vrije Universiteit Amsterdam, the Netherlands; ORCID: https://orcid.org/0000-0002-7824-7674*

[b]*University of Barcelona, Spain; ORCID: https://orcid.org/0000-0002-6755-1988*

* Marjolein K. Camphuijsen, Department of Educational and Family Studies, Van der Boechorststraat 7, 1081 BT, Amsterdam, the Netherlands. m.k.camphuijsen@vu.nl

Word count (incl. tables and references): 7,274

Teachers' beliefs about standardised testing and test-based accountability: comparing the perceptions and experiences of teachers in Chile and Norway

Abstract

The global popularity of test-based accountability appears to signal political trust in standardised assessments as valid and relevant measures of education quality. Nonetheless, research shows that educators' perceptions of standardised testing and test-based accountability can vary significantly, as do their responses to accountability demands. Considering the key influence of teachers' beliefs on the way in which they respond to education reforms, in this paper we examine teachers' beliefs and opinions about standardised tests and test-based accountability. We rely on a comparative study on the interpretations and experiences of standardised testing and test-based accountability demands of compulsory education teachers in Chile and Norway. These cases were selected following a most-different-systems design approach. By relying on data derived from an electronic survey (n=2,531) and in-depth interviews (n=60), the analysis shows how in both contexts, teachers are relatively critical about the validity, usefulness and fairness of the standardised tests, signalling a lack of trust of teachers in standardized testing and test-based accountability. Still, despite similar trends, some key differences in the beliefs of Chilean and Norwegian teachers are found, which highlight the influence of the sociocultural context in shaping teachers' beliefs. By illuminating how teachers in different contexts make sense of test-based accountability, our analysis contributes to the understanding of why the often-reported mismatch between policy expectations and policy outcomes might occur.

Keywords: accountability, standardized testing, policy enactment, teacher beliefs, Chile, Norway

Introduction

In recent decades, a growing number of countries have adopted large-scale standardised tests. Increasingly, these tests are used to measure the performance of schools and teachers, and to hold educational actors accountable. This reform approach, commonly referred to as test-based accountability, is often adopted or strengthened to ensure educational actors are responsive to and assume responsibility for centrally-defined learning goals and to promote data-driven decision-making (Verger, Parcerisa & Fontdevila, 2019). While the global popularity of test-based accountability appears to signal political trust in standardised assessments as valid, relevant and legitimate measures of education quality, research shows that educators' perceptions of standardised tests and test-based accountability vary significantly, as do responses to accountability demands and expectations around data-driven decision-making (e.g., see Jones & Egley, 2004; Holloway & Brass, 2018). Moreover, while often introduced or strengthened with the ambition of fostering school improvement, a number of studies, often conducted in high-stakes accountability contexts, have reported that schools may adopt practices that portray effectiveness and productivity while not actually making substantial improvements (Au, 2022).

Considering the often-reported mismatch between policy expectations and the ways in which test-based accountability policies are responded to in local contexts (McDermott, 2007), a significant body of research has focused on understanding how, why and under what circumstances test-based accountability instruments lead to improvements in educational practice. A number of scholars have emphasised that accountability instruments can only successfully change teacher practice when they influence teacher motivation (e.g., Hwa, 2021). Important in this regard is that teachers perceive the accountability instrument as 'sufficiently meaningful, legitimate or otherwise persuasive' (Hwa, 2021, p.237). Various studies corroborate the premise that teachers who view accountability instruments as legitimate are more likely to adapt and improve their educational practices (e.g., see Kim et al., 2019; Klinger & Rogers 2011). Various

researchers have also identified a number of mediating variables, such as school leadership styles and the school culture, which are likely to influence the reception of policy demands by teachers and hence their interpretation of and responses to these policy mandates (O'Day, 2002; Møller, 2009).

The above highlights the key influence of teachers' beliefs and perceptions on the way in which they respond to education reforms. As such, to gain a deeper understanding of the often-reported gap between policy expectations and policy outcomes (McDermott, 2007), it becomes crucial to get a better understanding of educators' varying beliefs. Against this background, in this paper, we aim to shed light on teachers' beliefs about and perceptions of standardised tests and test-based accountability, as well as of factors that could potentially explain these beliefs. Following a most-different-systems design approach, this paper reports on a comparative study on the interpretations and experiences of standardised testing and test-based accountability demands of compulsory education teachers in Chile and Norway. By relying on both quantitative and qualitative data, we show that teachers in both contexts are relatively critical about the validity and usefulness and fairness of large-scale standardised tests, the results of which are used for school accountability. We furthermore argue that uncritical interpretation of test scores by external audiences will prevent teachers from developing more positive views towards the tests and accountability system, while potentially eroding trust in teachers' work and professionalism.

Contextual background

Chile and Norway differ significantly from one another in a number of regards, including in terms of political institutional regimes, administrative traditions and levels of trust in public institutions, as portrayed in Table 1. Moreover, the two countries differ with regards to how test-based accountability systems have been designed.

Table 1

Country characteristics

Country	Chile	Norway
Welfare regime model	Liberal	Social-democratic
Politico-administrative tradition	New Public Management marketiser	Neo-Weberian
Dominant patterns of regulation of the teaching profession	Market and standards-based regulation	Professional knowledge & autonomy-based regulation
Societal trust in government and public institutions	Low	High
Trust in teachers	Lower percentage of teachers' feeling trusted by society	Higher percentage of teachers' feeling trusted by society

Source: adapted from Voisin and Dumay (2020), Verger, Parcerisa & Fontdevila (2019), and OECD (2020, 2022)

In the case of Chile, the country has undergone significant educational reforms since the 1980s, which have resulted in Chile having one of the most marketised education systems in the world. In Chile, schools are subject to "double accountability" (cf. Weinstein, Raczynski & Peña, 2020): market and administrative accountability. Moreover, the Chilean accountability system is characterised by high-stake consequences, which both teachers and schools face depending on, amongst other measures, the results of their students in the national standardised test, the well-known *Sistema de Medición de la Calidad de la Educación* or SIMCE (in English, School Quality Measurement System). The SIMCE test, which combines open and multiple-choice questions, is administered in grades 4, 6, 8 and 10 in reading, writing, numeracy, natural and social sciences[1]. This standardised test evaluates student achievement in a wide range of skills and contents in diverse areas and subjects of the national curriculum. Examples of accountability consequences include teachers' promotion decisions, individual and collective salary bonuses, reputational

[1] In Chile, standardized tests are not scored by teachers. Rather, the scoring is outsourced to private companies.

consequences (which affect parents' school choice), limitation of school autonomy, and the closure of schools that have been classified as underperforming for over a period of four years.

Inspired by a New Public Management governance logic, the so-called Preferential School Voucher Law (Law 20.248) and the Quality Assurance System (Law 20.529) have created new mechanisms, tools, and institutions (such as the Agency of Quality Assurance) to evaluate, classify and sanction low-performing schools. However, beyond the high stakes testing approach, the Quality Assurance System currently includes soft assessment and accountability tools such as qualitative reports from the school inspection, the assessment of the so-called 'Other Quality Indicators' (e.g., the school's socioemotional climate, parents' satisfaction with the school, etc.) and non-mandatory diagnostic assessments such as the 'Integral Learning Diagnostic' test, which is used as a self-evaluation tool to monitor schools' progress. In addition, the Agency also delivers external visits to low-performing schools to evaluate and support them in school improvement processes. These instruments are intended to foster quality improvement efforts and to promote the usage of data to inform both principals' and teachers' decision-making.

In the case of Chile, policymakers perceived a need to strengthen external accountability mechanisms to guarantee that both schools and teachers would behave in line with regulations and expectations around school improvement. Simultaneously, teachers' individual autonomy is limited in Chile and different investigations show that Chilean teachers experience a lack of trust in their professional judgement (Carrasco, 2013).

In the case of Norway, a National Quality Assessment System was introduced in 2004, which consists of various quality assessment measures, including national tests, mapping/screening tests, local tests for both summative and formative uses, international comparative achievement tests (e.g., PISA and PIRLS), Pupil Surveys, the School-Leaving Examination and the Craft Certificate (for an overview see Skedsmo, 2011). Many of these quality assessment measures serve a double purpose. On the one hand, they are meant to provide central authorities with information about the level of knowledge of Norwegian students, thereby

providing a basis for general decision-making as well as offering a means for central and local authorities to hold lower entities such as schools accountable. On the other hand, the measures are supposed to provide information to teachers, school leaders and local authorities, which can be used as a basis for quality improvement efforts. Local authorities are obliged to establish a system to follow up the results of quality assessment measures, and to prepare an annual report in which they assess the performance of primary and lower-secondary education in their jurisdiction and formulate strategies for improvement. National tests are among the prime measures used to hold teachers, schools and municipalities accountable for the extent to which their students meet national learning objectives. Currently, national tests, which consist of online multiple-choice assessments, are administered at the start of Grade 5, 7 and 8 in reading, numeracy and English[2] (Camphuijsen, Møller & Skedsmo, 2021). The Norwegian test-based accountability system relies on the publication of results (in a context of low levels of marketisation and restricted school choice) as well as follow-up by the local authority (the formal account-holder) as the prime accountability consequences.

Even though the Norwegian accountability system remains characterised by a relative lack of 'hard' consequences, and teachers' individual autonomy has been emphasised in reform efforts, it has been argued that the high levels of trust teachers traditionally enjoyed have been replaced by a situation wherein teachers increasingly are required to 'deserve' their trust. In this light, various studies report how Norwegian teachers perceived the introduction of test-based accountability as a sign of distrust in the teaching profession (Skedsmo & Mausethagen, 2016).

How teachers make sense of standardised testing and test-based accountability reforms

To shed light on teachers' beliefs about and perceptions of standardised tests and test-based accountability, policy enactment and sense-making theories form useful heuristic devices (Ball et

[2] In Norway, the scoring of the national tests is done by the computer, not by teachers.

al., 2011). These theoretical perspectives highlight the contentious and dialectical nature of policy enactment processes and emphasise how putting policy into practice involves individual and collective meaning making dynamics through which school actors decode external messages and new policy mandates. Meaning making processes do not occur in a vacuum, but rather take place within particular administrative and regulatory models that shape the teaching profession (Voisin & Dumay, 2020) and diverse micro-political organisations such as schools.

Thus, schools are key spaces where recently adopted education policies are shared and debated, and collective opinions and beliefs about policy are co-constructed. According to these theoretical perspectives, teachers are policy shapers who adapt external demands and policies to their worldviews and school contexts. From this point of view, teachers can actively appropriate, negotiate, reframe, and even resist new policy mandates. Subjective variables such as teachers' core beliefs, values and opinions act as a cognitive frame through which they filter, interpret and translate policy texts into everyday practices. These cognitive variables play a key role in mediating policy messages, and influence teachers' alignment with new policy programmes and instruments (Coburn, 2001). These analytical lenses help us to gain a fine-grained understanding of teachers' perceptions and beliefs about standardized testing and test-based accountability, as well as of the role of trust and legitimacy that standardized tests enjoy among various actors in explaining teachers' perceptions and experiences.

Data & methodology

In this study, we rely on both quantitative and qualitative data, collected in the context of a larger research project[3]. During the school years of 2018-2019 and 2019-2020, an online survey (see Levatino, 2021) was administered in a representative sample of primary and lower-secondary

[3] This study is part of the REFORMED research project, see: www.reformedproject.eu

schools in both countries[4]. In total, 1,225 teachers in Chile and 1,306 teachers in Norway completed our questionnaire. During the analysis of the quantitative data, we first carried out a contingency tables analysis. We used a nominal variable (country: 1. Norway and 2. Chile) as independent Xi and teachers' beliefs about the validity, the usefulness and fairness of the standardized test and test-based accountability as dependent ordinal variables (Yj) (see Table 2). To ascertain whether a difference existed in teachers' beliefs about standardised testing in the two countries, we conducted Pearson chi-square tests, which allowed to analyse the statistical significance for the observed relationships between independent and dependent variables. Finally, the strength of association between Xi and Yj was examined through a Cramer's V test.

In addition, upon administering the survey in both Chile and Norway, we carried out in-depth interviews with teachers in both countries. In doing so, we relied on a heterogeneous and purposive strategy sampling strategy and selected teachers with different personal characteristics (in terms of age, gender and years of work experience). The interviews were conducted between October 2018 and February 2020 and followed a semi-structured interview script[5], which was used in both contexts. Each interview was audio recorded and subsequently transcribed verbatim. In total, interviews were conducted with 28 teachers in Chile, working at 12 schools, and 13 teachers in Norway, working at 9 schools.

Table 2

Teacher beliefs variables and questions

Construct	Question wording	Answer options
Belief about the validity of the national test	A good teacher can be recognised by his/her students' results in national test	For each statement: strongly agree, agree, neither agree nor disagree, disagree, strongly disagree
	The results of national test do not adequately represent what students have learned and can do	

[4] The survey included questions about personal characteristics, teaching methods and classroom practices, the school context, interpretation and translation of standardized testing and test-based accountability demands, as well as job satisfaction and teacher efficacy (Levatino, 2021).

[5] The interview script included questions about beliefs about standardised testing and test-based accountability, data-use and pedagogic practices, teacher identity, autonomy and professionalism, and perceptions and experiences of interpersonal trust.

Belief about the usefulness of the national test	Preparation for national test takes too much time away from more important activities in school The content of national test tells us what the school's priorities are The results of national test do not provide useful information on student learning	For each statement: strongly agree, agree, neither agree nor disagree, disagree, strongly disagree
Belief about the fairness of the national test	To what extent do you consider it is fair… … to measure the quality of a school based on national test results? … to publicly disseminate national test results in the media and/or internet … that schools with different characteristics are compared on the basis of their national test results?	For each question: very fair, fair, unfair, very unfair

Source: adapted from Levatino (2021)

The analysis of the interview data consisted of three phases. First, we conducted a reading of all interview transcripts, while generating analytic memos. Second, we developed a codebook and coded all the interview scripts combining inductive and theory-driven codes that covered key themes such as teachers' opinions and beliefs about the validity, uselessness and fairness of the standardized tests, teachers' lived experiences of standardized testing and test-based accountability, pedagogic practices and data use, and teachers' perceptions on trust in standardized testing and in teachers. Third, we organised and analysed the codes by relying on qualitative content analysis.

Findings

Teachers' perceptions of the validity of the standardized test

Table 3 presents the findings from our electronic survey with regards to Chilean and Norwegian teachers' beliefs about the validity of the standardised tests in representing what students have

learnt and can do[6]. As can be derived from the table, in both Chile and Norway a majority of the respondents report that they (strongly) agree with the statement that standardised test results do not adequately represent what students have learnt and can do, while a minority of teachers (strongly) disagrees. Despite similar trends, results from the Chi-Square Test of Independence show that the relationship between the country and teachers' perceptions of the test's validity in representing what students have learnt and can do is statistically significant, X2(4, N=2,037) = 104.70, p = .000. The size of the difference, as measured by Cramer's V, is moderate, .23 (Cohen, 1988).

Table 3

Validity - The results of the national tests do not adequately represent what students have learned and can do

Country	Strongly disagree	Disagree	Neither agree nor disagree	Agree	Strongly agree	Total
Chile	54	104	184	398	377	1,117
	4.83%	9.31%	16.47%	35.63%	33.75%	100%
Norway	12	76	236	432	164	920
	1.30%	8.26%	25.65%	46.96%	7.83%	100%
Total	66	180	420	830	541	2,037
	3.24%	8.84%	20.62%	40.75%	26.56%	100%

Pearson chi2(4)= 104.7023 Pr=0.000
Cramer's V= 0.2267

The interview data provide further insight as to why some teachers question the validity of the standardised tests in measuring student learning. For example, in Norway, several teachers mention how the tests do not only measure how well a student can read, but also whether he/she is able to concentrate and sit still:

[6] In this particular question, the reference to 'what students have learnt and can do' is made to student learning in the competency/subject that is tested by the standardized test in question. This question does not refer to student learning in general or across the entire curriculum.

> What's a shame about those tests is that they also measure concentration and endurance. It is a test that takes 90 minutes. (...). They have to sit and work [for 90 minutes]. This can be difficult. And then you measure other things than just reading skills. - Lise[7]

In Chile, some of the interviewed teachers go even further in questioning the validity of the tests, as they wonder whether the tests measure students' learning at all:

> I think that [standardized tests like SIMCE] don't really measure learning, (…) because there are students and schools that have little familiarity with the instrument. - Laura

With regards to Chilean and Norwegian teachers' beliefs about whether a good teacher can be recognised by his/her students' results in the standardised test, Table 4 presents the findings from our survey. As portrayed in this table, a minority of the Chilean respondents report they (strongly) agree with the statement, while the majority of the Chilean respondents report they (strongly) disagree. In a similar vein, in the case of Norway, a minority of the respondents report that they (strongly) agree with the statement that a good teacher can be recognised by his/her students' standardised test scores, whereas a majority of the Norwegian teachers (strongly) disagrees with this statement. The results from the Chi-Square Test of Independence show that there exists a significant relationship between the country and teachers' perceptions of the validity of the test in reflecting teacher quality, $X2(4, N=2,034) = 48.57$, $p = .000$. Nonetheless, the size of the difference for this finding, as measured by Cramer's V, is low, .15 (Cohen, 1988).

Table 4

Validity - A good teacher can be recognized by his/her students' results in the national test

Country	Strongly disagree	Disagree	Neither agree nor disagree	Agree	Strongly agree	Total
Chile	368 32.95%	386 34.56%	215 19.25%	113 10.12%	35 3.13 %	1,117 100%

[7] To secure the anonymity of the respondents, all respondent names are pseudonyms.

Norway	283	371	220	38	7	919
	30.79%	40.37%	23.94%	4.13%	0.76 %	100%
Total	651	757	435	151	42	2,036
	31.97%	37.18%	21.37%	7.42%	2.06 %	100%

Pearson chi2(4)= 48.5753 Pr=0.000
Cramer's V= 0.1545

The interview data highlight that, in criticising the use of test results to measure the quality of individual teachers, many Chilean teachers express that the SIMCE test does not consider the complex conditions under which teachers' work is conducted in different school settings, which influence test results.

Other Chilean teachers, in contrast, are more positive about the validity of the standardised tests in reflecting teacher quality, simultaneously recognising the positive effect of high scores on the teacher's reputation. For example, Julieta interprets good performance as the logical consequence of the quality of teachers' work and their commitment to teaching:

> More than anything else, I think that it's because of the work you do; you commit yourself, you plan, you work [hard, and] the children learn. So, you feel that if the result is good, your name also stands out. – Julieta

With regards to the Norwegian data, the interviews confirm how most Norwegian teachers range from somewhat to very sceptical about the extent to which results represent the efforts and ability of individual teachers. One teacher mentions how results are always a collective responsibility:

> I know that other teachers at this school are very affected by the national tests, and when there has been a bad result, then it is not very nice (...), but there are so many teachers who have been, there are many teachers who are in a way "guilty", if you can call it that, because there are many teachers who have had the students over the years. But it is often the ones who had them last who will hear it the most... – Nina

Like some of the Chilean teachers, other Norwegian teachers go further in questioning who is responsible for test results, arguing that a range of different factors, including factors related to students' motivation or parental involvement, over which teachers do not have (full) control, play an important role in determining results:

> There is a limit to how much you can do yourself. There is also the children's own motivation, and the parents' own motivation… For students who do not have good results, this reason is perhaps almost the most important. If they are driven, it helps a lot. – Helene

Regardless of the teachers' acknowledgement that many factors influence learning outcomes, teachers in both Chile and Norway report they are the ones who are praised or blamed for performance.

Teachers' perceptions of the usefulness of the standardized test

Table 5 shows that a majority of the Chilean respondents report that they (strongly) agree with the statement that the results of the standardised test do not provide useful information on issues related to student learning, whereas a minority of Chilean teachers (strongly) disagrees with this statement. In contrast, in Norway, a minority of respondents report that they (strongly) agree with this statement, whereas 35% of the Norwegian respondents neither agree nor disagree and 42% of the Norwegian teachers (strongly) disagree with this statement. Results from the Chi-Square Test of Independence confirm that the relationship between the country and teachers' beliefs about the usefulness of the standardised test in providing information about student learning is significant, $X2(4, N=2{,}036) = 319.12$, $p = .000$. The size of the difference for this finding, as measured by Cramer's V, is medium-high, .40 (Cohen, 1988). It appears that Chilean teachers are less likely to perceive the tests as providing useful information about student learning compared to Norwegian teachers.

Table 5

Usefulness - The results of national tests do not provide useful information on student learning

Country	Strongly disagree	Disagree	Neither agree nor disagree	Agree	Strongly agree	Total
Chile	68	167	239	338	306	1,118
	6.08%	14.94%	21.38%	30.23%	27.37%	100%
Norway	42	344	318	174	40	918
	4.58%	37.47%	34.64%	18.95%	4.36%	100%
Total	110	511	557	512	346	2,036
	5.40%	25.10%	27.36%	25.15%	16.99 %	100%

Pearson chi2(4)= 319.1207 Pr=0.000
Cramer's V= 0.3959

The interview data provide insights as to how this finding could potentially be explained. For example, one group of Chilean teachers mentions that the national tests fail to provide important information which they would need to be able to make use of the test for data-driven decision-making. In this regard, the interviews reveal how some Chilean teachers consider metrics coming from private standardised tests[8], which some schools use, as more useful than the national test data. These teachers explain that the data coming from the private standardised tests is more comprehensive, since the private tests cover more areas and aspects, including students' socioemotional well-being. Moreover, the private tests also provide more detailed information, including individual student data, which allows teachers to see how students perform in each area or subject.

Another group of Chilean teachers recognise that some data from the national standardised tests might be useful, but explain that they rely mainly on their own professional expertise and judgment in order to identify students' needs, take pedagogical decisions and inform teaching practices.

[8] In Chile, various commercial providers offer private standardized tests to schools. Public and private subsidized schools in Chile receive funding from the State to contract external services from the school improvement industry, which include private standardized tests, teaching training, etc.

In contrast, interviews with Norwegian teachers highlight how most Norwegian teachers are (mildly) positive about the usefulness of the national tests in providing them with information about the students' learning. Nonetheless, also some Norwegian teachers explain to still miss important information:

> I think national tests have gotten better and better. I was not a big fan in the beginning. (...). There's still improvement to be made. There is still information that I miss, in particular at the class-level. We for example do not have the possibility to see what students answer when they answer a question wrong. – Rolf

Furthermore, as portrayed in Table 6, a minority of the Chilean respondents report that they (strongly) agree with the statement that the content of the standardised test tells them what the priorities of the school are/should be, whereas almost half of the Chilean teachers (strongly) disagrees with this statement. In the case of Norway, a minority of the respondents report that they (strongly) agree with the statement that the content of the standardised test tells them what the school's priorities are/should be, whereas almost half of the Norwegian respondents neither agree nor disagree and a little under half of the Norwegian respondents (strongly) disagree with this statement. Results from the Chi-Square Test of Independence show that the relationship between the country and teachers' perceptions on the usefulness of the standardised test in in telling what the school's priorities are/should be is significant, $X2(4, N=2{,}036) = 85.01$, $p = .000$. Nonetheless, the size of the difference as measured by Cramer's V is medium-low, .20 (Cohen, 1988).

Table 6

Usefulness - The content of national tests tells us what the school's priorities are

Country	Strongly disagree	Disagree	Neither agree nor disagree	Agree	Strongly agree	Total
Chile	200	305	322	204	86	1,117

	17.91%	27.31%	28.83%	18.26%	7.70%	100%
Norway	107	270	401	123	18	919
	11.64%	29.38%	43.63%	13.38%	1.96%	100%
Total	307	575	723	327	104	2,036
	15.08%	28.24%	35.51%	16.06%	5.11%	100%

Pearson chi2(4)= 85.0095 Pr=0.000
Cramer's V= 0.2043

Finally, the majority of Chilean respondents report that they (strongly) agree with the statement that the preparation for the standardised test takes too much time away from more important activities in school (as portrayed in Table 7), whereas a minority of Chilean teachers (strongly) disagrees with this statement. In contrast, in Norway, around one third of the respondents report that they (strongly) agree with the statement that the preparation for the standardised test takes too much time away from more important activities in school, whereas another third of the Norwegian teachers (33%) (strongly) disagrees with this statement. Also in this case, results from the Chi-Square Test of Independence show that the relationship is significant, X2(4, N=2,037) = 223.09, p = .000. The size of the difference for this finding as measured by Cramer's V is medium, .33 (Cohen, 1988). It appears that in Chile, teachers are more likely to have negative opinions about the effects of national tests on their work compared to teachers in Norway.

Table 7

Usefulness - Preparation for national tests take too much time away from more important activities in school

Country	Strongly disagree	Disagree	Neither agree nor disagree	Agree	Strongly agree	Total
Chile	55	119	252	382	310	1,118
	4.92%	10.64%	22.54%	34.17%	27.73%	100%
Norway	81	220	327	222	69	919
	8.81%	23.94%	35.58%	24.16%	7.51 %	100%
Total	136	339	579	604	379	2,037
	6.68%	16.64%	28.42%	29.65%	18.61 %	100%

Pearson chi2(4)= 223.0976 Pr=0.000
Cramer's V= 0.3309

Teachers' perceptions of the fairness of the standardised test

With regards to teachers' perceptions of the fairness of the standardised test, in both Chile and Norway, a minority of the respondents report to believe it is (very) fair to measure the quality of the school based on standardised test results, whereas the majority of respondents report to believe this is (very) unfair (as portrayed in Table 8). Despite similar trends, results from the Chi-Square Test of Independence shows there exists a significant relationship between the country and teachers' perceptions, X2(3, N=2,040) = 57.82, p = .000. Nonetheless, the size of the difference for this finding as measured by Cramer's V is low, .17 (Cohen, 1988).

Table 8

Fairness - To measure the quality of a school based on the national test results

Country	Very unfair	Unfair	Fair	Very fair	Total
Chile	523	407	159	31	1,120
	46.70%	36.34%	14.20%	2.77 %	100%
Norway	353	475	86	6	920
	38.37%	51.63%	9.35%	0.65 %	100%
Total	876	882	245	37	2,040
	42.94%	43.24%	12.01%	1.81%	100%

Pearson chi2(3)= 57.8244 *Pr=0.000*
Cramer's V= 0.1684

In addition, in both cases, a minority of the respondents report to believe it is (very) fair that schools with different characteristics are compared using standardised test scores, whereas the majority of respondents in both countries report to believe this is (very) unfair (see Table 9). Even so, results from the Chi-Square Test of Independence show there exists a significant relationship between the country and teachers' beliefs about the fairness of comparing schools, X2(3, N=2,040) = 57.82, p = .000. Nonetheless, the size of the difference as measured by Cramer's V, is low, .14 (Cohen, 1988).

Table 9

Fairness - that schools with different characteristics are compared on the basis of their national test results

Country	Very unfair	Unfair	Fair	Very fair	Total
Chile	627	366	95	31	1,119
	56.03%	32.71%	8.49%	2.77 %	100%
Norway	423	413	76	7	919
	46.03%	44.94%	8.27%	0.76 %	100%
Total	1,050	779	171	38	2,038
	51.52%	38.22%	8.39%	1.86%	100%

Pearson chi2(3)= 40.5019 *Pr=0.000*
Cramer's V= 0.1410

The interview data provide further insight as to why many teachers in both countries perceive school comparisons as unfair. That is, in both countries, teachers argue that results depend to a large extent on that year's student base. Consequently, as explained by one Norwegian teacher, 'if they were to say something about the actual effect of the schools, there would also have to be controls for socio-cultural background' (Helene). Moreover, in both cases, interviewed teachers explain that they feel that standardised test results are used to blame low-performing schools, irrespective of the work and effort put in by the school staff.

Nonetheless, regardless of their critical attitude towards school comparisons, Table 10 shows how almost half of the Chilean respondents report to believe it is (very) fair to publicly disseminate the standardised test scores in the media or on the internet, whereas a little over half of the respondents report to believe this is (very) unfair. In contrast, in Norway, a minority of the respondents report to believe it is (very) fair to publicly disseminate the standardised test scores in the media or on the internet, whereas the majority reports to believe this is (very) unfair. Results from the Chi-Square Test of Independence confirms that the relationship is significant, X2(3, N=2,036) = 154.45, p = .000. Nonetheless, the size of the difference for this finding, as measured by Cramer's V, is medium-low, .28 (Cohen, 1988).

Table 10

Fairness - To publicly disseminate national test results in the media and/or in the internet

Country	Very unfair	Unfair	Fair	Very fair	Total
Chile	316	316	410	75	1,119
	28.29%	28.29%	36.71%	6.71%	100%
Norway	321	425	163	10	919
	34.93%	46.25%	17.74%	1.09%	100%
Total	637	741	573	85	2,036
	31.29%	36.39%	28.14%	4.17 %	100%

Pearson chi2(3)= 154.4572 *Pr=0.000*
Cramer's V= 0.2754

The interview data illuminate some of the reasons behind diverging beliefs. That is, some Chilean teachers express a positive attitude towards the publication of results as they explain to value transparency. Other Chilean teachers, who express a more critical attitude towards the dissemination of results, explain this by pointing towards the performative effects and competition between schools that the dissemination of test results promote.

In the Norwegian case, the interview data reveal that Norwegian teachers in particular are critical with regards to how media actors use and disseminate test results, arguing that media coverage is 'not very nuanced' and contributes to the image of 'winners' and 'losers', which is felt as unfair. One teacher recalls:

> The worst thing was that the results came in the newspaper, and we were hung out in last place, the worst in the whole of [name of municipality] and that… I remember that feeling, it was so [exhaling loudly] ... We felt that we worked in the worst school, but we worked maybe most of all, but no one saw everything we did. I think it was so unfair. - Anette

Moreover, a number of Norwegian teachers mentioned that the significant attention paid to the results throughout the year is problematic, as it seems to result in a situation where some schools (excessively) prepare the students for the tests. As one teacher explains:

> I think more teachers would have been positive about national tests if principals had managed to convey national tests as more than just a test, but also as an opportunity to make changes in teaching, an opportunity to take action *after* the national tests, not in advance. - Andreas

Discussion and conclusions

In this paper, we have reported on a comparative study of teachers' beliefs about and experiences of standardised testing and test-based accountability in Chile and Norway. Our investigation shows how in both contexts, teachers are relatively critical about the validity, usefulness and fairness of the standardised test, signalling a lack of teacher trust in standardised testing and test-based accountability. That is, a majority of Chilean and Norwegian teachers consider that the standardized tests do not adequately represent what students have learnt and can do and form a poor descriptors of the quality of their work. Moreover, our analysis shows that the majority of teachers in both contexts perceive it as unfair to measure the quality of a school based on standardized test scores and to compare schools with different characteristics using test scores.

Still, despite similar trends, some key differences in the perceptions of Chilean and Norwegian teachers are found. More specifically, with regards to the perceived validity and usefulness of the standardized tests, Chilean teachers appear more likely to perceive the tests as an invalid measure of what students have learnt and can do, and as providing little useful information about student learning. This latter finding might relate to the fact that national standardized test scores in Chile omit important details which teachers would need to use the tests to inform their teaching practices. Moreover, it seems that Chilean teachers are more likely to hold a negative opinion about the effects of standardized testing on their work. On the other hand, Norwegian teachers seem more likely to express a critical attitude towards the public dissemination of test results.

At first glance, the almost equally critical attitude of Chilean and Norwegian teachers towards standardised testing and test-based accountability, and the even more critical attitude of

Norwegian teachers towards the dissemination of test results, might seem counterintuitive. In the Norwegian case, teachers face few high-stakes consequences based on their students' performance, while Chilean teachers face significant gains and losses. One possible explanation for the (more) critical attitude of Norwegian teachers might be the lack of compatibility between the accountability system and Norwegian teachers' notions of who is to be trusted. That is, as recently argued by Hwa (2021, p.244), compatibility between teacher accountability and generalised notions as to who is to be trusted can 'help to legitimize these instruments in teachers' eyes, which facilitates the influence of the accountability instruments over teacher motivation and teacher practice'. Whereas the accountability system in Norway might be compatible with the notions of politicians or citizens as to who is to be trusted, the lack of alignment with teachers' own notions as to who is to be trusted might contribute to the failure to positively influence Norwegian teachers' beliefs and motivation.

In the case of Chile, the more positive perceptions of teachers towards the market uses of standardised tests and test-based accountability might be explained by cultural changes deriving from the market reforms initiated in the late 1980s and the policies' long trajectory and consolidation. That is, after decades of profound market reforms, some market values and principles such as transparency and school choice might be internalised into principals' and teachers' rationalities (Falabella, 2020). As a consequence, market uses of test-based accountability might enjoy higher legitimacy among teachers. In both cases, this would imply the sociocultural context (Hwa, 2021) plays a key role in shaping teachers' beliefs about standardised testing and test-based accountability.

In addition, what seems to play a role in shaping the critical attitude of both Chilean and Norwegian teachers is the trust and legitimacy that standardized tests enjoy among key external audiences. In both contexts, teachers argue that actors outside of the school, such as local and national authorities, parents and media outlets, often take test scores at face value and as telling an important truth about teacher or school quality, while teachers strongly disagree with the notion

that test scores adequately reflect their abilities and efforts. Considering that assessment experts have shown that no single test can measure learning across an entire curriculum and many factors (beyond the teacher's role) affect learning outcomes, it is problematic that national test scores sometimes become interpreted as proxies of education and teacher quality. Literature in the field of the sociology of quantification offers fruitful explanations as to why performance indicators such as standardised tests are often perceived as objective, reliable and robust measures. In particular, the social process of commensuration, which implies 'the comparison of different entities according to a common metric' by turning qualities into numbers seems crucial to understand the power of performance metrics, and the legitimacy they enjoy among external audiences (Espeland & Stevens, 1998, p.314. This in part because numbers are often more valued by people due to their ease of comparison and wider held beliefs about the objectiveness of numbers.

It has been suggested that usage of multiple student assessments might reduce such narrow interpretations of education quality, while simultaneously lowering the risk of practices such as teaching to the test or curriculum narrowing. Existing research indeed underlines the importance of the design of the assessment framework in order to promote trust in test results as well as to prevent inappropriate practices (OECD, 2013). At the same time, a better understanding among key external audiences of what assessment data can and cannot show seems to form another important condition for teachers to develop a more positive view towards the tests and the accountability system. In other words, promoting assessment literacy among external audiences, such as national and local authorities as well as parents, can be an important way to ensure trust in the system. This is also important considering that uncritical interpretation of the scores might erode societal trust in teachers' work and professionalism (Daliri-Ngametua, Hardy & Creagh, 2021).

In addition to external audiences, building capacity and promoting assessment literacy among school leaders and teachers also seems important to foster improvement of educational

practices. At the school and classroom level, test results might identify gaps in student learning or reveal areas where further school-level attention is needed. A good understanding among school actors of what test data can and cannot tell, as well as the ability to diagnose the causes of low performance, and the capacity to formulate improvement strategies, can therefore promote effective usage of test results for school improvement purposes. With this in mind, one way of increasing the legitimacy of test-based accountability systems in teachers' eyes might be to hold teachers accountable for 'making the most productive uses of the resources available to them in an effort to move toward the goal' (Leithwood & Earl, 2000, p.5), instead of holding them uniquely or primarily accountable for students' achievement in external assessments.

To conclude, our investigation highlights that many Chilean and Norwegian teachers perceive standardised testing and test-based accountability as a contentious and controversial policy option. Considering the key influence of teachers' beliefs on how they respond to education reforms, our analysis contributes to the understanding of why the often-reported mismatch between policy expectations and policy outcomes might occur. Future research could explore the mediating role of school leadership on how teachers perceive and use test results and examine the impact of varying teacher beliefs on how they respond to accountability expectations.

Data availability statement

The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

Funding

This work was supported by the European Research Council under Grant 680172.

References

Au, W. (2022). *Unequal by design: High-stakes testing and the standardization of inequality*. Routledge.

Ball, S. J., Maguire, M., & Braun, A. (2011). *How schools do policy: Policy enactments in secondary schools*. Routledge. https://doi.org/10.4324/9780203153185

Camphuijsen, M. K., Møller, J., & Skedsmo, G. (2021). Test-based accountability in the Norwegian context: exploring drivers, expectations and strategies. *Journal of Education Policy, 36*(5), 624-642. https://doi.org/10.1080/02680939.2020.1739337

Carrasco, A. (2013). Mecanismos performativos de la institucionalidad educativa en Chile: pasos hacia un nuevo sujeto cultural. *Observatorio cultural*, 15(1), 4-10.

Coburn, C. E. (2001). Collective sensemaking about reading: How teachers mediate reading policy in their professional communities. *Educational evaluation and policy analysis*, *23*(2), 145-170.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum.

Daliri-Ngametua, R., Hardy, I., & Creagh, S. (2021). Data, performativity and the erosion of trust in teachers. *Cambridge Journal of Education*, 1-17. https://doi.org/10.1080/0305764x.2021.2002811

Espeland, W. N., & Stevens, M. L. (1998). Commensuration as a social process. *Annual review of sociology*, 24(1), 313-343. https://doi.org/10.1146/annurev.soc.24.1.313

Falabella, A. (2020). The ethics of competition: accountability policy enactment in Chilean schools' everyday life. *Journal of Education Policy*, 35(1), 23-45. https://doi.org/10.1080/02680939.2019.1635272

Holloway, J., & Brass, J. (2018). Making accountable teachers: The terrors and pleasures of performativity. *Journal of Education Policy*, *33*(3), 361-382. https://doi.org/10.1080/02680939.2017.1372636

Hwa, Y. (2021). Contrasting Approaches, Comparable Efficacy? How Macro-Level Trust Influences Teacher Accountability in Finland and Singapore. M. Ehren & J. Baxter (Eds.), *Trust, Accountability and Capacity in Education System Reform: Global Perspectives in*

Comparative Education, edited by Melanie Ehren and Jacqueline Baxter, (pp. 222–251). New York: Routledge.

Jones, B. D., & Egley, R. J. (2004). Voices from the Frontlines: Teachers' Perceptions of High-Stakes Testing. *Education policy analysis archives*, *12*(39), 1-34. https://doi.org/10.14507/epaa.v12n39.2004

Kim, J., Sun, M., & Youngs, P. (2019). Developing the "will": The relationship between teachers' perceived policy legitimacy and instructional improvement. *Teachers College Record*, *121*(3), 1-44. https://doi.org/10.1177/016146811912100301

Klinger, D. A., & Rogers, W. T. (2011). Teachers' perceptions of large-scale assessment programs within low-stakes accountability frameworks. *International journal of testing*, *11*(2), 122-143. https://doi.org/10.1080/15305058.2011.552748

Leithwood, K., & Earl, L. 2000. Educational accountability effects: An international perspective. *Peabody Journal of Education*, *75*(4), 1-18. https://doi.org/10.1207/S15327930PJE7504_1

Levatino, A. (2021). *Surveying Principals and Teachers: Methodological insights into the design of the REFORMED Questionnaires*. REFORMED methodological papers No. 2.

McDermott, K. A. (2007). "Expanding the moral community" or "blaming the victim"? The politics of state education accountability policy. *American Educational Research Journal*, *44*(1), 77-111. https://doi.org/10.3102/0002831206299010

Møller, J. (2009). School leadership in an age of accountability: Tensions between managerial and professional accountability. *Journal of Educational change*, 10(1), 37-46. https://doi.org/10.1007/s10833-008-9078-6

O'Day, J. (2002). Complexity, accountability, and school improvement. *Harvard educational review*, 72(3), 293-329. https://doi.org/10.17763/haer.72.3.021q742t8182h238

OECD (2013). *Synergies for Better Learning: An International Perspective on Evaluation and Assessment*. Paris: OECD Publishing.

OECD. (2020). *TALIS 2018 Results (Volume II) Teachers and School Leaders as Valued Professionals*. Paris: OECD Publishing.

OECD (2022). *Trust in government (indicator)*. Paris: OECD. doi: 10.1787/1de9675e-en (Accessed on 18 April 2022).

Skedsmo, G. (2011). Formulation and realization of evaluation policy: inconsistencies and problematic issues. *Educational Assessment, Evaluation and Accountability, 23*(5), 5-20.

Skedsmo, G. & Mausethagen, S. (2016). Emerging accountability policies and practices in education: The case of Norway. In J. Easley & P. Tulowitzki (Eds.), *Educational Accountability. International perspectives on challenges and possibilities for school leadership* (pp. 205–223). Routledge

Verger, A., Parcerisa, L., & Fontdevila, C. (2019). The growth and spread of national assessments and testbased accountabilities: a political sociology of global education reforms. *Educational Review, 71*(1), 5–30.

Voisin, A., & Dumay, X. (2020). How do educational systems regulate the teaching profession and teachers' work? A typological approach to institutional foundations and models of regulation. *Teaching and Teacher Education*, *96*, 1-16. https://doi.org/10.1016/j.tate.2020.103144

Weinstein, J., Raczynski, D., & Peña, J. (2020). Relational trust and positional power between school principals and teachers in Chile: A study of primary schools. *Educational Management Administration & Leadership*, 48(1), 64-81. https://doi.org/10.1177/1741143218792912