Article

Educational Policy
1–30

DOI: 10.1177/08959048221142048
journals.sagepub.com/home/epx
SAGE

Understanding the Stakes: The Influence of Accountability Policy Options on Teachers' Responses

Antonina Levatino[1], Lluís Parcerisa[2], and Antoni Verger[1]

Abstract

Under test-based accountability, side-effects —including practices to inflate test results, often seen as cheating—are usually associated to so-called high-stakes policies. However, the influence of different types of stakes in the generation of this type of practices has been overlooked in education research. Based on a survey experiment, our results indicate that the type and level of stakes of accountability systems (e.g., high- vs. low-stakes, material vs. symbolic) do not differ in triggering side-effects. Counterintuitively, individual symbolic consequences trigger similar reactions among teachers than material incentives. In-depth interviews give insights into the social mechanisms that lead to symbolic effects having such an influence in understanding teachers' reactivity to accountability.

Keywords

test-based accountability, standardized tests, accountability policy options, undesired responses, teacher reputation, side-effects, cheating

[1]Universitat Autònoma de Barcelona, Cerdanyola, Spain
[2]Universitat de Barcelona, Barcelona, Spain

Corresponding Author:
Antonina Levatino, Department of Sociology, Universitat Autònoma de Barcelona, Edifici B - Campus, Bellaterra, Barcelona, Cerdanyola 08193, Spain.
Email: Antonina.levatino@uab.cat

Introduction

In the last two decades, test-based accountability (TBA) policies have acquired much centrality in the governance of educational systems all over the world. This form of accountability consists on the evaluation of student performance in different areas of knowledge through large-scale assessments and standardized tests. In countries with TBA in place, school actors, including teachers and school principals, face consequences of a different nature according to their level of adhesion to centrally defined learning standards and academic achievement of their students (Amrein-Beardsley & Holloway, 2019; Lingard et al., 2013).

TBA has generated passionate educational debates for different reasons, which include its uncertain and inconclusive effects on quality and equity. A recent OECD Education working paper on TBA and student achievement suggests that in some countries (especially low- and middle-income) TBA could have positive effects on students' performance and detrimental impact on equity due to the increase of the learning gap (Torres, 2021). In a similar vein, some investigations consider that, in certain contexts, TBA can improve learning outcomes (Chiang, 2009), while others point out that these policies rather act as triggers of undesired responses at the school and/or classroom level that jeopardize educational inclusion and the breadth of the curriculum (Berliner, 2011; Booher-Jennings, 2005; Ohemeng & McCall-Thomas, 2013). By undesired responses we refer to a wide range of actions and reactions performed by educational actors in an attempt of raising test scores, without this necessarily entailing a substantive improvement in educational quality (Murnane & Cohen, 1986). Examples of teacher undesired responses range from the intensive practice of teaching to the item and narrowing the curriculum to more illicit actions, such as cheating during the test administration by suggesting or directly modifying students' responses to test items (Jacob & Levitt, 2003) or altering the pool of the students tested through the exclusion of underperforming students (Hamilton et al., 2002; Hofflinger & von Hippel, 2020).

Undesired responses to external testing, especially those involving illicit actions, have been usually reported in countries with high-stakes accountability systems, such as the US, England, Australia, and Chile (Jones et al., 2003; Koretz, 2017; Martinelli et al., 2018; Nichols & Berliner, 2005). These are systems that involve formal material consequences (including teachers' promotion and income decisions, or school intervention in case of poor performance) that educational authorities put in place to sanction or reward account givers (Au, 2007). In countries with high-stakes TBA, cases of school cheating have even transcended academic circles and attracted media attention

(Amrein-Beardsley & Berliner, 2002; El Mercurio, 2004; Hofflinger & von Hippel, 2020; Nichols & Berliner, 2005).

In contrast, in so-called low-stakes educational accountability regimes (e.g., France, Germany, and Norway), the consequences of TBA are considered "softer" or of a more "symbolic" nature (Maroy & Voisin, 2017; Thiel & Bellmann, 2017). Consequences are symbolic in the sense that are normally associated with the visibility given to test results, but not with material rewards and/or sanctions for account givers (Maroy & Pons, 2019). Nonetheless, as we develop below, recent research argues that low-stakes accountability also has the potential to generate performance pressure on schools and teachers as well (Figueiredo et al., 2016; Hammersley-Fletcher et al., 2021; Maaranen & Wågsås Afdal, 2020). This seems to be linked to the implications the public diffusion of test results may have for the definition of what a "good teacher" and a "good school" are (Camphuijsen, 2020; Maroy & Pons, 2019). It is also acknowledged that, in marketized educational contexts, that is, where there is school choice freedom and school funding depends on the number of enrolled students, symbolic consequences may also be perceived as "harder" to the extent they can influence future school demand and, accordingly, school resources (Maroy & Voisin, 2017). Indeed, the complex interaction between accountability and other types of educational policies—such as school choice, school browsers, and/or enrollment policies—is something that, in real education settings, contributes to problematize "low-stakes" and "high-stakes" as watertight categories.

Given that undesired responses corrupt the relationship between performance indicators and actual performance (Van Thiel & Leeuw, 2002), which is at the basis of the whole TBA idea, it is important to deepen our understanding of whether the different types of consequences of test results have the potential to generate these kinds of responses. Finding out about the effects of different TBA policy options presents, however, a number of challenges for the researcher. Different consequences often coexist within the same accountability system and, when comparing different policy options, many contextual factors can act as confounders. This hampers the possibility to estimate the isolated effect of a given consequence at stake by using observational data.

Existing academic literature on undesired effects of TBA has mainly focused on the investigation of different types of strategic practices to "game the system" (e.g., Amrein-Beardsley et al., 2010; Booher-Jennings, 2005) and on the detection of cheating practices with the aim of estimating their widespread adoption and exploring their contextual determinants (e.g., Ehren & Swanborn, 2012; Ferrer-Esteban, 2013; Hibel & Penn, 2020; Jacob & Levitt, 2003). Some recent cross-country comparative analyses on TBA have

investigated broader accountability effects on teachers' practices and decisions, but without delving into the relationship between different accountability policy options and undesired responses (Houtsonen et al., 2010; Lennert da Silva & Mølstad, 2020; Osborn, 2006). Furthermore, existing studies normally rely on self-reported data through survey or interview techniques. These techniques are not only insufficient to estimate the isolated effect of different forms of consequences in TBA systems, but are also subject to social desirability bias in the reporting of actors' own undesired behavior (Amrein-Beardsley et al., 2010). To our knowledge, no study has explored yet the extent to which social desirability bias can be at stake when asking school actors about these kinds of practices.

In this study, we apply first a randomized survey experiment conducted among a sample of 1,130 Chilean teachers, where we ask them to rate the likelihood of a third person and of themselves cheating after having presented a situation where we alter the consequence at stake. The goals of the experiment are (1) to estimate the effects that policy options (in our case, different types of consequences attached to TBA) may have on teachers' responses and (2) to explore whether self-reporting on undesired behavior may be subject to social desirability bias. After analyzing the experimental results, we conduct a qualitative analysis of 22 interviews among teachers focused on the performance pressure they feel, with the aim of (3) deepening the understanding about the social motives of the experimental results obtained.

The research has been carried out in Chile since this country has undergone structural educational reforms since the 1980s that have promoted, among other changes, accountability policies based on external evaluations and a liberalized school choice system (Bellei, 2015). Nowadays, Chile is one of the countries in the world that, for a longer period of time, has experienced an accountability policy based on external evaluations. This system has been indeed viewed as a mean to (1) promote school competition via the publication of test results in a totally open school choice system (Bellei, 2015) where schools receive funding according to the number of students enrolled; (2) assure educational quality and (3) held schools, principals, and teachers accountable via a Quality Assurance System and merit-based pay policies that aim to incentivize school improvement and control schools (Falabella, 2020; Mizala & Schneider, 2014). In the Chilean context, the problem of the undesired effects of TBA is currently the subject of public, political, and academic debate, something that makes the generation of new evidence about the causal mechanisms behind school responses to accountability policy more pressing (e.g., Falabella, 2020; Hofflinger & von Hippel, 2020; Pino et al., 2016).

Our research does not aim to estimate to what extent cheating is widespread among Chilean teachers, but rather, to estimate to what extent different kinds of accountability policies are viewed by teachers as more likely to induce an illicit behavior, complementing this analysis with the exploration of how teachers experience TBA performance pressures.

In the paper, before presenting our findings and discussing them, we first revise the main theoretical explanations that could enlighten the emergence of side-effects in TBA contexts and, second, present our methodological approach.

The Undesired Effects of TBA Policy: Theoretical Underpinnings

TBA policies strongly rely on goal-setting theories, according to which the setting up of specific objectives and the association of the achievement of such objectives to incentive schemes positively impact on employees' performance (Locke, 1968). In particular, setting clear at the same time that challenging objectives would lead "to even higher performance than [simply] urging people to do their best" (Latham & Locke, 2007, p. 291).

The translation of goal-setting premises in educational policy is not straightforward. Research has shown that, while for simple tasks, incentives can be effective and increase productivity, for more complex processes that involve multiple tasks, the worker might narrow focus and energy on the specific tasks that are incentivized, and neglect the others (Gibbons, 1998). Experimental studies have also demonstrated that, when confronted by challenging performance goals, subjects tend to overstate their productivity more than when confronted with the "do your best" stimulus (Welsh & Ordóñez, 2014). Furthermore, the achievement of too ambitious goals and/or excessively high stakes may conduce to opportunistic behavior (Madaus & Clarke, 2001; Patrick et al., 2018).

The corruption of performance indicators was predicted many years ago, when Lindquist (1951, pp. 152–153) pointed out that:

> Because of the nature and potency of rewards and penalties associated in actual practice with high and low achievement test score of students, the behavior measured by a widely used test tends in itself to become the real objective of instruction, to the neglect of the (different) behavior with which the ultimate objective is concerned.

Later, the well-known Campbell's law similarly suggested that "the more any quantitative social indicator is used for social decision-making, the more

subject it will be to corruption pressures and the more apt it will be to distort and corrupt the social processes it is intended to monitor" (Campbell, 1979, p. 85). TBA policies are based on the theoretical assumption that sanctions and/or rewards will increase school actors' motivation to perform well. Nonetheless, as noted by Amrein-Beardsley et al. (2010, p. 5), a high pressure to perform might induce a person "to engage in practices that ordinarily are not typical of that person." As performance pressure tends to be conceived as the main driver of cheating in public accountability systems (Patrick et al., 2018), it has been generally assumed that the higher the stakes, the higher the likelihood that school actors will engage in non-desired behaviors such as cheating (e.g., Ferrer-Esteban, 2013; Nichols & Berliner, 2005). According to Nichols and Berliner (2005), cheating would be even "inevitable" when the stakes are high.

From this it derives that the nature of the stakes is an important feature in the definition of TBA policies. The literature on accountability in education tends to distinguish between high- and low-stakes TBA, depending on the type and the intensity of the consequences attached to the test results (Verger & Parcerisa, 2017; Maroy & Voisin, 2017). For a long time, it has been implicitly assumed that material consequences (e.g., economic rewards, school intervention, teachers' tenure and promotion decisions, etc.) are mainly present in high-stakes TBA systems, whereas low-stakes accountability involves mainly symbolic consequences, that is, making test results visible to different actors (including authorities public, families, other colleagues, the media, etc.). Nonetheless, it has been argued that symbolic consequences can also raise the stakes, especially where the educational context is marketized (Maroy & Voisin, 2017). Making test scores public can have tangible repercussion on school demand and enrollment and, accordingly, shape the organizational and educational actions of schools (Figlio & Loeb, 2011). Not coincidentally, the undesired responses of school actors is often reported in contexts where market dynamics and high-stakes testing combine (Amrein-Beardsley et al., 2010; Au, 2011; Jones et al., 2003; Koretz, 2017). However, the assumption that material consequences is what induce undesired behaviour cannot be taken for granted. High-stakes TBA normally involve both material and symbolic consequences, and it is almost impossible to disentangle which type of consequence is a major driver of this type of behavior.

In fact, emerging research conducted in lower-stakes accountability policy environments has also documented TBA exerting pressure on schools and teachers, and altering their decisions and practices (Feniger et al., 2015). These findings highlight the weight symbolic consequences can have in driving school actors' decisions, even where no market dynamics are in place.

Actually, as both high- and low-stakes TBA systems share the existence of symbolic consequences, evidence of non-desired effects in both types of policy settings can lead one to hypothesize that concerns attached to the visibility of test results might actually be a prominent source of performance pressure with the potential of leading to undesired responses as well. The mechanism of commensuration, that is, the fact that the qualities of good education and good teaching are equalized to learning metrics and related performance indicators (see Espeland, 2013), would play a key explanatory role.

Existing educational research has shown that TBA policies are related to the emergence of side effects, and even to cheating. In the news, as well as occasionally in academic research, teachers' cheating is usually conceptualized as a "moral failure" and teachers incurring in these practices are portrayed as "rotten apples" (Jacob & Levitt, 2003), instead of trying to understand whether and how certain policy features and institutional characteristics are inducing people to behave this way. The harsh denouncement of this kind of practices in the public sphere make it plausible to consider that self-reporting is subject to social desirability bias, so that people directly asked about their involvement in cheating practices by an external observer might tend to underreport it.

In summary, conducting research on the side effects of TBA is challenging for at least three main reasons. First, in high-stakes accountability systems, both material and symbolic consequences coexist; it is not thus possible to disentangle the role played by the different incentives in affecting actors' responses and, in fact, many contextual factors can act as confounders. Second, in accountability systems where only symbolic consequences exist, it is not possible to know whether undesired behavior would be more pervasive if material consequences were also at stake. Third, certain responses to accountability pressure such as those involving cheating tend to be underreported. To overcome these methodological challenges, we have designed a randomized survey experiment that tries to isolate the effect that consequences attached to TBA may have on teachers' responses. By means of in-depth qualitative interviews, we then deepen on what lies behind the results obtained with the experimental approach.

Methodology

The methodology of this study is made up of multiple components. First, we use a factorial survey experiment (1) to assess the isolated effect that policy design options (in our case, different types of consequences attached to TBA) may have on teachers' reactivity, and (2) to explore the extent social

desirability bias can be at stake when individuals are directly asked to report about illicit behavior in response to performance pressure. It is important to underline that, to pursue this goal, we use the estimated likelihood that, according to the respondents, a given policy option could derive in an undesired behavior. We consider it as a good proxy of the pressure a policy option exerts over teachers. Afterward, we use 22 in-depth qualitative interviews conducted among teachers (3) to deepen our understanding of the experimental results and explore teachers' experienced pressure related to performance and standardized tests.

The Experiment

Design of the vignettes. The experiment used is a vignette, a factorial survey experiment where each participant is treated with a hypothetical situation that randomly presents different characteristics (cf. Auspurg & Hinz, 2015), that is, experiment conditions/treatments (Atzmüller & Steiner, 2010); in our case, the consequences attached to test scores. The use of an experimental approach presents several advantages for the achievement of our research goals. In contrast with observational studies, because of the absence of correlations between the different manipulated treatments, as well as between the treatments and the respondents' characteristics, experiments allow the unbiased estimation of the isolated causal effects of each manipulated treatment (Druckman & Leeper, 2012; Leeper, 2018). In addition, due to the complexity of the facts presented and their presentation as hypothetical situations in a vignette, it is less likely that the answers will be subject to a social desirability bias (Steiner et al., 2016).

In the framework of our experiment, each participant, randomly assigned to one of five groups, reads a vignette that describes the situation of a teacher whom a colleague advises to cheat in the context of a new national assessment to obtain better results in the test. Specifically, the colleague advises the teacher to send the underperforming students to the school library to do an alternative activity on the day of the test. Aimed at altering the pool of tested students, this behavior constitutes what is often seen as an illicit practice. The vignette is manipulated with regard to the possible consequences that the test may have if students obtain poor results. As can be seen in Table 1, apart from the control group that read a vignette where no consequence is mentioned, we designed the treatments according to two dimensions: the type of the consequences (material or symbolic) and the locus of accountability (the school or the individual teacher). Every treatment occurred with the same frequency.

Table 1. Experiment Treatments.

Group	Type of consequence	Locus of accountability	Treatment
Baseline condition (Control group)	—	—	—
Treatment 1	Material	School	School funding reduction
Treatment 2	Material	Teacher	Salary bonus for individual teachers
Treatment 3	Symbolic	School	School reputation affected
Treatment 4	Symbolic	Teacher	Teacher reputation affected

The gender of the protagonist of the vignette and the social composition of the neighborhood of the school where the teacher teaches (middle class or vulnerable) are also randomly manipulated. This ensures that these elements are not left to the imagination of the respondent and prevents the results from being influenced by these factors. At the end of the vignette, the respondent rates from 1 (Not at all likely) to 7 (Extremely likely) both the likelihood of the protagonist of the vignette cheating and the likelihood of doing so if the respondent personally were in the same situation.[1]

In a first step, to verify whether social desirability plays a role, we compare the answers to the question asked in third person ("likelihood of the protagonist of the vignette engaging in the behavior") with those given to the question made in first person ("likelihood of the respondent engaging in the behavior if she/he were in the same situation"). To check whether differences are statistically significant, as the variables of interest are ordinal, we perform a Wilcoxon Signed Rank test (Wilcoxon, 1945). This is a nonparametric test used to compare two paired samples when assumptions of the paired *t*-test are not satisfied or when the variable of interest is ordinal. In a second step, we estimate the effects of our treatments on the likelihood of the protagonist of the vignette engaging in the described illicit practice. Given that the dependent variable has seven categories placed in an order, to determine to what extent the experimental conditions are predictors of the dependent variable, we carry out ordinal logistic regressions. To facilitate the interpretation of the results, we show proportional odds ratios. For the sake of robustness, we also conduct a Kruskal-Wallis test to determine if the reported likelihood was different for the five groups. A post-hoc Dunn test with Holm p-adjustments is then conducted to identify pairwise differences between groups.

External validity. One of the main criticisms to studies based on vignettes concerns the fact that both the situations presented and the follow-up questions merely concern hypothetical scenarios, and that actual behavior is not observed. Nonetheless, studies conducted to contrast vignette judgments with actual behavior bolster confidence in the use of this method by showing how at least the direction of the effect, as well as the strengths of the treatments, usually match (e.g., Hainmueller et al., 2015). In our experiment, we estimate the isolated effect that each accountability consequence has on the extent to which teachers perceive a given illicit behavior more or less likely to occur. This constitutes an important proxy of the performance pressure leading to actual cheating that each of the consequences exerts on teachers.

To enhance the experiment's external validity, we apply the experiment to real teachers (cf. Auspurg & Hinz, 2015). The underlying data were collected through a survey conducted among 1,130 teachers from a sample of 81 schools in the three largest urban areas of Chile during school years 2018/2019 and 2019/2020.[2] We chose to conduct the research in Chile because it constitutes a high-stakes TBA system and, as underlined above, in such systems, both types of consequences (material and symbolic) exist. We can therefore be assured that the experiment treatments were perceived as plausible and realistic by the participants.

The participants were selected through a two-step stratified sampling strategy. In the first step, 200 schools were selected through a systematic probability proportional to size sampling strategy. All the schools in the sample impart primary or lower secondary education (Basic Education).[3] The explicit strata were constructed according to the type of ownership, while within each stratum, schools were sorted by the implicit variables: province, municipality, and school size.[4] Within each selected school, a sample of 24 teachers was drawn from two sampling strata according to whether or not the teacher was currently teaching a subject tested in the national standardized test (SIMCE). The average number of teachers who responded per school was almost 15.[5] The questionnaire was self-administered through the Qualtrics online platform. Table 2 shows the characteristics of the respondents and of the schools where they are working.

In our sample, public schools are overrepresented (+8.2 percentage points), whereas independent private schools are underrepresented (–8.6). The percentage of subsidized private schools in the dataset is in line with the initial sample. Regarding the school location, all the Santiago provinces except North Santiago are underrepresented. In contrast, schools in Concepción, Valparaiso and North Santiago are overrepresented. The average size of schools in our data is slightly below the levels of the initial sample (26.3 vs. 30.4). Not all the schools in the initial sample participated in the study; therefore we cannot consider the obtained sample

Table 2. Respondent and School Characteristics.

	M (*SD*)/%	Range
Respondent characteristics (*n*. 1,130)		
Gender		1–3
Female	74.58	
Male	25.24	
Other gender	0.18	
Ever prepared students for SIMCE		1–4
Currently preparing	62.53	
Not currently preparing, but prepared less than 3 years ago	14.17	
Not currently preparing, but prepared more than 3 years ago	13.20	
Never prepared	10.10	
School characteristics (*n*. 81)		
School ownership		1–3
Public	43.21	
Subsidized private	51.85	
Independent private	4.94	
School SES quartile*		1–4
First	33.33	
Second	29.63	
Third	22.22	
Fourth	14.81	
School location		1–3
Concepción	33.33	
Metropolitan Region of Santiago	49.50	
Valparaíso	17.28	

*The quartiles have been calculated by inverting the School Vulnerability Index (IVE).

a probability sample. However, a representative sample is not needed to test the causal effects of the experiment in an accurate way (cf. Auspurg & Hinz, 2015; Mullinix et al., 2015). The use of inferential statistics with experimental design where respondents are recruited from a non-probability sample is well established. In experimental research, the aim is to detect the probability that the effect seen in the specific subjects participating in the experiment was due to the treatment and not to any other confounding variable (Lang, 1996). Mullinix et al. (2015) have recently compared results obtained from convenience samples and population-based samples. The substantial similarities found further strengthen confidence in the usefulness of non-representative samples for experimental research.

Internal validity checks. In experiments, the unbiased treatment effects' estimation is enabled by two characteristics of the experimental design, the maximum zero-correlation of the vignette dimensions (orthogonality) and the maximum variance of the levels (balance), which guarantee internal validity (cf. Kuhfeld et al., 1994). To check for successful randomization and variation, and assess the quality of the experimental data, we calculate bivariate Pearson's correlations between the treatments and respondents' characteristics and between all the treatments (Table 3). The correlations are all near zero, demonstrating that the core principles of experiments have been successfully applied.

Qualitative Interviews

The qualitative phase is based on semi-structured interviews (Wengraf, 2001) with a total of 22 Chilean teachers (see Table 4). Our aim is to delve (Brinkmann & Kvale, 2018) into teachers' rationalities and subjectivities in order to better understand how they live and experience TBA pressures. We consider this research technique crucial to better understand the reasons why teachers would see illicit behavior more likely to occur under specific TBA policy options. The interview script has different modules, such as teachers' biographical information, opinions and perceptions about TBA and market pressures, data use, teaching practices and professional autonomy. In this study, we focus on the part of the interviews centered on teachers' general perceptions of TBA pressures and their daily lived pressures. During the interviews, the teachers typically spontaneously linked their experiences of pressure with the stakes attached to the test, without being explicitly asked by the interviewer.

The analysis aims at deepening our understanding of the results of the experiment, with particular regard to individual reputational concerns and the reasons behind them. The analysis followed an iterative process and combined deductive and emergent codes. After coding the entire interviews, we extracted all the quotes about teachers' perceptions and experiences of TBA pressures. Once the quotes were organized, the three authors read the transcripts and added analytic memos (Saldaña, 2021).

The selection of the participants followed a purposive and heterogeneous strategy (Schreier, 2018). Given the focus of the study, we selected teachers who work in primary or lower secondary education (Basic Education) and teach (or have recently taught) in grades and subjects assessed in the SIMCE test. The sample is heterogeneous in terms of the characteristics of the participants (years of experience, age, gender) and of the school where they are working (school ownership, socioeconomic composition, performance category).

Table 3. Correlation Matrixes (Internal Validity Checks).

Correlation between vignette dimensions and respondents' characteristics

		Respondents' characteristics			Characteristics of respondents' schools				
Vignette dimensions		Age	Gender	Ever prepared students for SIMCE	Ownership	Average SIMCE performance	SES quartile	Location	School identifier
	Consequences	.0231	–.0440	–.0176	–.0138	–.0192	.0233	–.0420	–.0115
	Gender	.0045	.0157	–.0375	.0240	.0267	–.0188	.0022	.0454
	Social composition	–.0116	–.0066	–.0182	.0130	.0050	.0177	–.0303	–.0285

Correlations between vignette dimensions

	Consequences	Gender	Social composition
Consequences	1.000		
Gender	–.0006	1.000	
Social composition	–.0019	.0018	1.000

Table 4. Sample of Participants in the Qualitative Scrutiny.

School	Ownership	School composition (SES)	Performance category	No. of teachers interviewed
1	Public	Low SES	Low-Medium	1
2	Private	High SES	High	2
3	Private—subsidized	High SES	Medium	2
4	Public	Med-high SES	High	3
5	Public	Low SES	Insufficient	2
6	Private—subsidized	Med-high SES	Low-Medium	2
7	Private—subsidized	Med-high SES	Low-Medium	2
8	Private	High SES	High	1
9	Public	Med-high SES	Medium	1
10	Private—subsidized	Low-med SES	Low-Medium	2
11	Private—subsidized	Low SES	Medium	2
12	Public	Low-med SES	Medium	2
Total				22

Findings

Results of the Experiment

In order to analyze whether self-reporting on what we call undesired responses might be subject to social desirability bias, we check for differences in the answers obtained to the question on the likelihood of the protagonist of the vignette aiming to alter test results by avoiding underperforming students taking the test, and those obtained when asking directly to the respondents. Figure 1 displays the frequencies of the answers given to the questions asked in third person with those of the question in first person. It shows how, for the question asked about the protagonist of the vignette, answers spread across the categories (*median = 4; mean = 3.58; SD: 1.96*), but for the direct question, there was a clear polarization of a large number of responses in the category "Not at all" (*median = 1; mean = 2.09; SD: 1.64*).

A more detailed look at the individual differences between the answers referring to the protagonist of the vignette and to the direct question (self-reporting) reveals that 35.93% of the respondents gave the same answer to the two questions, 58.58% reported a higher likelihood of the protagonist of the vignette cheating compared with their own likelihood of doing the same, and only 5.49% reported a higher personal likelihood of engaging in the illicit practice described in the vignette. The results of the Wilcoxon Signed Rank

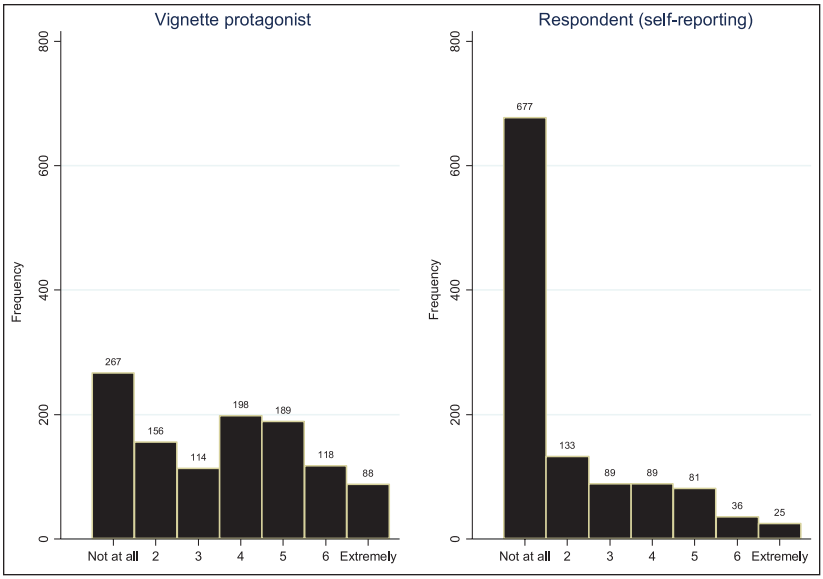

Figure 1. Reported likelihood of performing the illicit behavior (comparison: vignette protagonist vs. respondent).

test, statistically significant at the .01% level (–22.358, *p*-value: .0000), clearly indicate that respondents tend to rate higher the likelihood of undesired behavior when they are referring to a third person than when referring to themselves. This suggests that social desirability can be a serious issue that can alter the results of research based on direct questioning and/or self-reporting. Thus, research based on conventional data-gathering techniques may suffer from the social desirability effect and underestimate the spread of non-desired behavior.

In order to analyze the effects of the treatments on the likelihood of cheating, in the vignette, we use as a dependent variable the answers referring to the protagonist of the vignette, as they appear to be less subject to social desirability bias. Four models have been calculated: Model 1 contains only the experimental treatments whereas Models 2 and 3, respectively, also include, as controls, variables concerning respondents' school characteristics and individual characteristics. In Model 4, variables of both levels are included, while in Model 5 school dummies are also added. Tabular presentation of results can be found in Table 5.[6] For the sake of transparency, when discussing the results, we also report 95% confidence intervals (CI) to also

Table 5. Effects of the Treatments on the Likelihood of Vignette Protagonist Engaging in the Illicit Practice (Odds Ratios).

	1	2	3	4	5
Treatments					
Consequence (ref. No consequences)					
School funding reduction	1.355	1.328	1.189*	1.363	1.403*
	(0.22)	(0.22)	(0.26)	(0.22)	(0.24)
Individual salary bonus	1.522**	1.546**	1.570**	1.596**	1.730**
	(0.25)	(0.26)	(0.26)	(0.27)	(0.30)
School reputation affected	1.563**	1.577**	1.609**	1.626**	1.687**
	(0.26)	(0.26)	(0.26)	(0.27)	(0.29)
Individual reputation affected	1.701***	1.732***	1.757***	1.790***	1.842***
	(0.28)	(0.29)	(0.29)	(0.30)	(0.32)
Gender of the vignette protagonist (ref. Male)					
Female	0.836	0.836	0.852	0.939	0.862
	(0.09)	(0.09)	(0.09)	(0.10)	(0.09
Social composition vignette protagonist's school (ref. Vulnerable)					
Middle class	0.929	0.941	0.927	0.939	0.968
	(0.10)	(0.10)	(0.10)	(0.10)	(0.11)
Controls					
Gender of the respondent	NO	NO	YES	YES	YES
Respondent has ever prepared students for SIMCE	NO	NO	YES	YES	YES
Respondent's school ownership	NO	YES	NO	YES	YES
Respondent's school SES quartile	NO	YES	NO	YES	YES
Respondent's school location	NO	YES	NO	YES	YES
School dummies	NO	NO	NO	NO	YES
N. obs.	1,130	1,130	1,128	1,128	1,128

Note. Standard-errors in parentheses.
*$p < .05$. **$p < .01$. ***$p < .001$.

give information about the uncertainty of the estimations (Hinkle et al., 2003; Zientek et al., 2012).

The results of Model 4 show how, compared to the reference category, having some consequences associated with the test increases the probability of teachers rating the likelihood of cheating higher. Interestingly, the results remain inconclusive for the consequence "school funding reduction," a material consequence normally considered high stake. The results for the other three consequences are clearer regarding the direction of the effect, but

uncertain regarding its strength. The effect of the treatment regarding individual reputation displays the highest values and is statistically significant at the .01% level (est. 1.842; IC: 1.316–2.579; *p*-value: .000), compared with those regarding individual salary bonus (est. 1.730; IC: 1.231–2.431; *p*-value: .002) and school reputation (est. 1.687; IC: 1.207–2.357; *p*-value: .002).[7] A Kruskal-Wallis test, conducted for the sake of robustness, confirms these results. It shows that there is a statistically significant difference in the variable "likelihood of the protagonist of the vignette to cheat" between the five groups ($\chi^2(4) = 13.267$, $p = .01$). Nevertheless, a pairwise post-hoc Dunn test with Holm p-adjustments is only significant for the differences concerning the pairs: "individual salary bonus" versus the control group ($p = .037$), "school reputation affected" versus the control group ($p = .023$), and "individual reputation affected" versus the control group ($p = .005$). These results challenge the common assumption that accountability instruments are more likely to generate undesired effects when attached to material consequences. It is true that, in a marketized school context such as the Chilean one, school reputation can be also considered high stakes for its (material) impact on school demand and enrollment (Figlio & Loeb, 2011; Maroy & Voisin, 2017). Nonetheless, compared to the other treatments, it is individual reputation the consequence whose results are the ones that show a higher statistical significance. This is interesting considering that this is the consequence that can be considered a "pure" symbolic one (cf. Maroy & Pons, 2019). It is therefore relevant to explore more in depth teachers' rationalities and experiences with TBA, with a focus on why individual reputation seems to matter so remarkably.

Teachers Under Accountability Pressure: Why Reputational Consequences Matter

From the interviews, it appears that the visibility of test results generates a performative environment in schools that increases the lived pressure of teachers. References to individual reputational consequences emerged very frequently during the interviews, and were often associated to the performance pressure teachers experience:

> **Teacher:** Yes, I may feel under pressure. I may think of the results. I would like that the results would be good.
> **Interviewer:** And, why?
> **Teacher:** I think that, above all, for the work you do: you commit, you plan, you work so that the children learn. So, you feel that, if the results are good, your name also stays up. [Laugh]. (Vanny, language teacher).

The response of this language teacher, specializing in one the subjects that the national assessment covers, reflects how crucial it is to obtain good results for individual teachers. Prominent performance metrics allow teachers to build and consolidate a positive reputation as a good professional among their peers and regulate good impressions of themselves. Obtaining good scores on external assessments plays an important role for receiving recognition from their "significant others," including the school staff and the management team.

Nonetheless, in line with what Booher-Jennings (2005, p. 252) argues, the preoccupation with test results seems to go beyond a "hedonist" willingness to impress. Rather, TBA is a source of affliction and preoccupation, in which the subjective effects of test results' visibility cannot be disentangled from the material conditions of teachers' work. In a context like the Chilean one, where early career teachers often do not hold a permanent contract and teacher turnover is high, reputational concerns can be occasionally related to the probability of renewing their contract or even losing their job, as explained by this math teacher:

> This year I do not have SIMCE, thank God. The other teacher has it. But it is so stressful! Apart that it is a whole process that. . . I have a temporary contract here, so you can realize, it is my fourth year working in this school and I still have a temporary contract. Next year they can fire me, they can go without my services. . . (Victor, math teacher).

Furthermore, the publication of test results leads to a situation where students' level of achievement is informally used to question teachers' work and commitment, as a language teacher points out:

> I think it [the publication of results] is an issue because, after that, comes the questioning of the teacher, for instance, to the teacher who prepares [the students for the] test. And I believe that. . .there, they [public educational authorities] do not take the time to see these other factors that intervene in a test of this nature, so, well, you have - as a good Chilean says - to 'bite the bullet[8]': Yes, I receive all the criticism, but deep down I highly value my work, so, then, if a good or bad job was done, I commit because, no matter how my students did it, I will support them anyway, no matter of what will happen with me. . . (Fernanda, language teacher).

In line with previous studies (e.g., Camphuijsen, 2020; Harris & Graham, 2019), our qualitative data show that TBA instruments have the power to transform and redefine existing shared understandings of "good teaching." Specifically, students' performance on standardized tests becomes a widely

accepted proxy of teacher and school quality, which is something that exacerbates test-related pressures over teachers. The following two quotations reflect this idea well.

> I think [the pressure and questioning over teachers work] comes from the fact that there is also this other part that a good result is a good teacher, and it makes us start to judge the others for the work they do, for their [teaching] practices. (Fernanda, language teacher).

> Here or in other schools where I have been working, it has been like having a 'stone in the chest' or a 'gun pointed at the head', like it is said, 'Results must be increased, results must be increased'; but clearly, it is because they measure us, so you cannot afford to decrease these results. Always, as parents (including myself as a mother, as an empowered actor), one asks as the first thing: 'What results do you have at SIMCE?', so it [test results] is something that one always analyzes, that is always present in your mind. (Mónica, math teacher).

Expressions like the ones used by these interviewees illustrate the negative emotions that many teachers experience when exposed to the risk of "receiving all the criticism." These expressions give an accurate description of the feelings of blameworthiness that account-giving actors experience when exposed to blame in their professional context. As another teacher states:

> Even though it [getting good test results] is not something that keeps me awake, to say it in some way, they [test results] are issues that are discussed with parents, issues that are also addressed within the groups, or inside me, by the people with whom I share, they talk about SIMCE. (Teresa, math teacher).

In TBA frameworks, symbolic consequences act as "solidary incentives" (see Clark & Wilson, 1961), as they are related to the desire of the individuals to hold membership in a community. To the extent the exposure to blame can be experienced as a threat to the membership to the "significant" community, it might enhance a need for self-protection and self-preservation, as well as a greater disposition toward illicit practices (Mitchell et al., 2018). Falabella (2020, pp. 12–13) points out that TBA policies involve a "sticky web of multidirectional surveillance and pressure to make them [teachers] feel accountable, motivated, and committed," as well as "to share responsibility, blame, and feelings of guilt." TBA policies are complex because they trigger "multiple accountability pressures" (Aleksovska & Schillemans, 2020, p. 2) that not only come from formal educational authorities, but also from other relevant audiences such as school principals, owners, and/or families. This language teacher highlights this idea:

> And yes, it generates pressure. It generates pressure to work for SIMCE. Having said that, it also will depend on how you focus your work as a teacher and take SIMCE in hand, but, yes, it implies more work, you have to give account to the UTP [pedagogical coordinator, in English] you have to give account to the coordinator, you have to give account to the students themselves, to your peers, to this external evaluator. . . (Gustavo, language teacher).

The pressures generated by TBA policies can be perceived as both external and internal to the individual. On the one hand, performance metrics appear as an objective indicator or benchmark, which allows the public administration, the school management, and parents to judge teachers' work and project performance expectations. On the other hand, these new policy technologies re-culturalize the teaching profession in such a way that teachers internalize entrepreneurial values and beliefs associated with the culture of performativity (Ball, 2003). This way, teachers may conceive performance pressure as endogenously generated or as triggered by their own professional ethos, as this quotation suggests:

> In some way, you also feel that one is also being evaluated. So, if it [the test] went bad for them (the students), it went also bad for me. It is not only their [the students'] results. It is not a matter of dignity, it is not that, all in all, one wants that it [the test] goes well. So, I feel, that one self-stresses oneself, that one self-imposes pressure on oneself. (Natalia, language teacher).

In the following quote, a math teacher illustrates the so-called "terrors" and "pleasures" of performativity, which previous research has associated with TBA policies as well (see Ball, 2003; Holloway & Brass, 2018).

> **Interviewer:** Ok and how did you experience all this [referring to the TBA policy]?
> **Teacher:** Badly. . .Regardless of whether one is for or against the standardized test, everyone likes [to obtain good results] and that students learn and take a decent test, right?
> **Interviewer:** Well, you were saying [that] you lived it badly. Why?
> **Teacher:** Because. . .they are also measuring teachers' work, this is the first part, and second, it is not funny to see bad results, it is not something that gives happiness, because in a certain way, it [the test] measures what the students can get to know. So, if the students are well evaluated, one comes out with [higher] professional job satisfaction, and the school also does well, and I also feel that students are doing even better. (Gabriel, math teacher).

These words also point out the extent to which school actors internalize that performance metrics are becoming synonym with teachers' and schools'

quality (Sullivan et al., 2020). In the same way that excessive performance pressure can affect teachers' experience negatively, obtaining good results in the standardized test can lead to both individual and collective satisfaction. Thus, beyond material incentives, "purposive incentives," that is, incentives centered on the pleasure coming from the achievement of valued goals (Clark & Wilson, 1961), have also a motivating effect for teachers.

However, despite performance metrics re-signify the notion of "teacher quality," internally, teachers also retain their own beliefs about what it means to be a good teacher. This creates a tension between conceptions of teacher quality that derive from, on the one hand, managerial approaches to outcomes-based education and, on the other, the pedagogic discourse. As a language teacher notes:

> But the relevance given to the SIMCE. . .I think it categorizes you as a teacher, as a good teacher, actually, do you understand me. . .? But to be a good teacher, you do not only have to have good results in SIMCE, but you have to, I say it again. . .you have to listen, you have to advise, you have to design other things. . . (Mario, language teacher).

This qualitative analysis illustrates the primary role of individual reputation in TBA frameworks that the survey experiment highlighted. The reputational effects of TBA need to be understood in the context of the centrality acquired by these tests in the governance of education, but also as the result of complex phenomena such as educational quality being commensurate with standardized tests results. Performance metrics simplify and, in turn, make visible, comparable, and legible teaching work and quality for multiple audiences (Holloway & Brass, 2018, p. 379). Regardless of the material incentives and sanctions attached to test results, and beyond the concerns that a non-consolidated professional status can generate among early-career teachers, the very visibility of test results entails high performative pressure for teachers on its own. In many cases, TBA exposes teachers to a situation in which they feel permanently scrutinized and judged by a wide range of actors and audiences. The judgment and questioning coming from colleagues and peers becomes a particular matter of concern, since teachers conceive test results as a device that can alter their status within their professional community.

Conclusions

TBA policies have become central in the governance of educational systems. These policies are increasingly used to monitor schools' outcomes, evaluate

teachers' work, and promote instructional improvement. However, at the same time, multiple investigations have documented the emergence of adverse or side effects of TBA, and how these can undermine the quality and equity of education. One of the effects that have aroused an intense academic and, in some contexts, public debate has been the spread of test cheating practices among school actors as a shortcut to increase test results (Ferrer-Esteban, 2013; Jacob & Levitt, 2003). Existing literature tends to link the emergence of such practices to high-stakes accountability. However, these dynamics have recently been documented in lower-stakes TBA policy environments as well (Feniger et al., 2015).

By means of a survey experiment, the present study has analyzed first the contentious relationship between accountability policies and teachers' responses by paying attention to the changing effect of different policy options. Specifically, the experiment has allowed us to understand to what extent certain accountability policy stimuli dispose teachers to (over-)react by avoiding underperforming students taking the test. Our results show how large-scale assessments with consequences associated to the results dispose teachers to inflate test results, but that the nature of the stakes do not significantly differ in triggering this type of practices. In fact, and counterintuitively, according to our findings, individual reputational consequences, which are normally considered as purely symbolic and "low-stakes," constitute the consequence whose effect shows the highest coefficient and statistical significance.

In a second stage, by means of qualitative interviews, we have deepened the understanding of the mechanisms that make individual reputational concerns increase the likelihood of engaging in illicit practices. We interviewed teachers of tested subjects and asked them about their lived experiences of pressure coming from TBA. In line with the findings of our experiment, in the interviews, reputational concerns emerged frequently, and with more assiduity than concerns with material or economic consequences. Our qualitative evidence indicates the centrality of the mechanism of commensuration in explaining the role of reputational consequences of TBA. To a great extent, performance metrics put pressure on teachers indirectly by redefining teaching quality and professionalism. Furthermore, the very fact of being monitored through external assessments predisposes teachers to aim to control the image they portray, even if this involves some levels of "inauthentic fabrication" of assessment results (see Webb, 2006).

Our study reveals that, more than driven by a mere desire to appear or improve the own position, reputational concerns are often accompanied by affliction, preoccupation, and fear of losing membership and/or credibility within the professional community. Symbolic consequences act as solidary

incentives (see Clark & Wilson, 1961), as they are related to the desire of the individuals to hold membership in a community and reputation vis-à-vis significant others (Finnigan & Gross, 2007). It should be added that, particularly in an educational context like the Chilean one—where teachers' working conditions are often precarious and the percentage of teachers with permanent contracts is relatively low (see Toledo Figueroa & Wittemberg, 2015)—material and symbolic consequences are not sealed compartments. Teachers concerns with the impact of low-test results on their credibility and reputation are difficult to disentangle from professional career concerns, including job continuity issues, in a market-oriented educational system where dynamics of school and teacher competition are widespread.

In sum, our results confirm that professional reputation plays a central role in the trajectory and impact of TBA policies. Our research indeed shows that the effects of accountability policies on teachers' practices are not only explained by formal consequences but also, and especially, by informal and symbolic ones. Reputational concerns might instill a sense of threat in teachers and the need for self-protection and self-preservation which is something that, as has been found by behavioral scientists (e.g., Mitchell et al., 2018), can lead to practices that oversize work achievements. Our findings highlight that the symbolic consequences of accountability, far from harmless, "low-stakes" or "soft," are experienced as meaningful by teachers. This finding challenges the common assumption that TBA policies predominantly generate undesired effects when attached to material consequences. The results thus encourage future scholarship to critically scrutinize the strict differentiation between high- and low-stakes categories in explaining the varying effects of accountability policy in educational practice. Rather, we suggest that it would be more appropriate to understand the stakes as a continuum in which different types of consequences interact and feedback each other.

Additionally, TBA policies create new power dynamics in education and engender a dense network of hybrid accountability relationships (horizontal and vertical) made up of multiple audiences that could put pressure on teachers, beyond official account holders (see Benish & Mattei, 2020). Our study suggests the need to further disentangle these relationships to elucidate the sources of performative pressures that affect teachers' daily lives. A policy implication deriving from our findings is that assessment frameworks, to be meaningful as a formative instrument, cannot give so much centrality to test data in the way they assess teachers' work. Whereas at the level of research, our findings suggest that studies on accountability, assessment and education need to go beyond the analysis of formal consequences and take into account how school actors "live" external evaluations and its related stakes—including the reputational and symbolic ones.

At a more methodological level, the experiment results confirm that reporting illicit behavior is subject to a social desirability bias, as respondents reported a greater likelihood of occurrence of this type of behavior when referring to a third person than when answering about themselves. This has important research implications, because although other studies have found that some people report having cheated when directly asked (Amrein-Beardsley et al., 2010), due to social desirability, these types of practices might have been underestimated in previous studies. Our study confirms that an experimental approach can contribute to investigate socially sensitive phenomena in education (such as cheating in external evaluations, but also issues of student discrimination and exclusion by some school providers), which are difficult to capture using conventional research designs.

This research is not exempt from shortcomings and future research is needed to address them. First, even though the experimental results give information about the direction of the isolated effects of three of the consequences explored, because of sample size limits they do not allow us to determine in a precise way the strength of these effects. Second, even though our qualitative analysis gives clear insights into the importance of individual reputation, it could be that the precariousness of the Chilean teacher working force overstates the importance of individual reputational effects, for the reasons stated above. Future investigation replicating this mixed-methods approach with a larger sample of teachers and in different country settings is therefore urgently needed. This research could unravel, among other things, how teachers who operate in educational systems with less-prevalent market competition experience and react to TBA stakes, and analyze to what extent their responses differ to those observed in Chile.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the European Research Council under Grant 680172 (Reformed project).

ORCID iDs

Antonina Levatino https://orcid.org/0000-0001-7245-3592

Lluís Parcerisa https://orcid.org/0000-0002-6755-1988

Notes

1. The vignette text and the two follow-up questions can be consulted in Levatino (2021).
2. More information on the survey content can be found in Levatino (2021).
3. According to the International Standard Classification of Education (ISCED), in Chile, Basic Education includes ISCED 1 and 2. For this reason, we selected teachers who work in both primary (from grades 1 to 6 and ages 6 to 11) and lower secondary education (grades 7–8 and ages 12–14) (UNESCO, 2012).
4. More information on the school sampling strategy, on data coverage and their representativeness can be found in Ferrer-Esteban (2022).
5. In schools with fewer than 24 teachers, the survey was applied to all the teaching staff. More information on the teacher sampling strategy, on data coverage and representativeness can be found in Ferrer-Esteban (2022).
6. As the variables "gender" and "ever prepared students for SIMCE" have one missing value each, results of Models 2, 4, and 5 concern 1,128 observations.
7. An odds ratio of 1.842 means, for example, that respondents who have seen the treatment regarding individual reputation have 1.842 times the odds of having answered that is extremely likely versus the combined other lower categories, compared to the control group.
8. This is the authors' own translation of the Spanish expression "poner el pecho a las balas," which means to accept and/or to face something unpleasant.

References

Aleksovska, M., & Schillemans, T. (2020) Dissecting multiple accountabilities: A problem of multiple forums or of conflicting demands? *Public Administration*, *100*(3), 711–736.

Amrein-Beardsley, A., & Berliner, D. (2002). *An analysis of some unintended and negative consequences of high-stakes testing*. Education Policy Studies Laboratory, Arizona State University. Retrieved August 5, 2021, from https://nepc.colorado.edu/sites/default/files/EPSL-0211-125-EPRU.pdf

Amrein-Beardsley, A., Berliner, D. C., & Rideau, S. (2010). Cheating in the first, second, and third degree: Educators' responses to high-stakes testing. *Education Policy Analysis Archives*, *18*(14), 1–36.

Amrein-Beardsley, A., & Holloway, J. (2019). Value-added models for teacher evaluation and accountability: Commonsense assumptions. *Educational Policy*, *33*(3), 516–542.

Atzmüller, C., & Steiner, P. M. (2010). Experimental vignette studies in survey research. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, *6*(3), 128–138.

Au, W. (2007). High-stakes testing and curricular control: A qualitative metasynthesis. *Educational Researcher*, *36*(5), 258–267.

Au, W. (2011). Teaching under the new Taylorism: High-stakes testing and the standardization of the 21st century curriculum. *Journal of Curriculum Studies*, *43*(1), 25–45.

Auspurg, K., & Hinz, T. (2015). *Factorial survey experiments*. SAGE.

Ball, S. J. (2003). The teacher's soul and the terrors of performativity. *Journal of Education Policy*, *18*(2), 215–228.

Bellei, C. (2015). *El gran experimento: Mercado y privatización de la educación chilena*. Lom ediciones.

Benish, A., & Mattei, P. (2020). Accountability and hybridity in welfare governance. *Public Administration*, *98*(2), 281–290.

Berliner, D. (2011). Rational responses to high-stakes testing: The case of curriculum narrowing and the harm that follows. *Cambridge Journal of Education*, *41*(3), 287–302.

Booher-Jennings, J. (2005). Below the bubble: "Educational triage" and the Texas accountability system. *American Educational Research Journal*, *42*(2), 231–268.

Brinkmann, S., & Kvale, S. (2018). *Doing interviews*. SAGE.

Campbell, D. T. (1979). Assessing the impact of planned social change. *Evaluation and Program Planning*, *2*(1), 67–90.

Camphuijsen, M. (2020). Coping with performance expectations: Towards a deeper understanding of variation in school principals' responses to accountability demands. *Educational Assessment, Evaluation and Accountability*, *33*, 427–453.

Chiang, H. (2009). How accountability pressure on failing schools affects student achievement. *Journal of Public Economics*, *93*(9), 1045–1057.

Clark, P. B., & Wilson, J. Q. (1961). Incentive systems: A theory of organizations. *Administrative Science Quarterly*, *6*(2), 129–166.

Druckman, J. N., & Leeper, T. (2012). Learning more from political communication experiments: Pretreatment and its effects. *American Journal of Political Sciences*, *56*(4), 875–896.

Ehren, M., & Swanborn, M. (2012). Strategic data use of schools in accountability systems. *School Effectiveness and School Improvement: An International Journal of Research, Policy and Practice*, *23*(2), 257–280.

El Mercurio (2004). Escándalo: Descubren a alumnos con "torpedo" en la prueba SIMCE. https://www.emol.com/noticias/nacional/2004/11/10/163647/escandalo-descubren-a-alumnos-con-torpedo-en-prueba-simce.html (accessed 29 July 2021)

Espeland, W. (2013). Commensuration and cognition. In K. A. Cerulo (Ed.), *Culture in mind* (pp. 63–88). Routledge.

Falabella, A. (2020). The ethics of competition: Accountability policy enactment in Chilean schools' everyday life. *Journal of Education Policy*, *35*(1), 23–45.

Feniger, Y., Israeli, M., & Yehuda, S. (2015). The power of numbers: The adoption and consequences of national low-stakes standardised tests in Israel. *Globalisation, Societies and Education*, *14*(2), 183–202.

Ferrer-Esteban, G. (2013). *Rationale and incentives for cheating in the standardized tests of the Italian assessment system* (Working Paper No. 50). Fondazione Agnelli.

Ferrer-Esteban, G. (2022). Sampling Strategy. REFORMED Methodological Papers No 3.

Figlio, D., & Loeb, S. (2011). School accountability. In E. A. Hanushek, S. Machin, & L. Woessman (Eds.), *Economics of education* (pp. 383–421). Elsevier-Handbooks in Economics.

Figueiredo, C., Leite, C., & Fernandes, P. (2016). The curriculum in school external evaluation frameworks in Portugal and England. *Research in Comparative and International Education*, *11*(3), 282–297.

Finnigan, K. S., & Gross, B. (2007). Do accountability policy sanctions influence teacher motivation? Lessons from Chicago's low performing schools. *American Educational Research Journal*, *44*(3), 594–629.

Gibbons, R. (1998). Incentives in organizations. *Journal of Economic Perspectives*, *12*(4), 115–132.

Hainmueller, J., Hangartner, D., & Yamamoto, T. (2015). Validating vignette and conjoint survey experiments against real-world behavior. *Proceedings of the National Academy of Sciences*, *112*(8), 2395–2400.

Hamilton, L. S., Stecher, B. M., & Klein, S. P. (2002). *Making sense of test-based accountability in education*. Rand Corporation.

Hammersley-Fletcher, L., Kılıçoğlu, D., & Kılıçoğlu, G. (2021). Does autonomy exist? Comparing the autonomy of teachers and senior leaders in England and Turkey. *Oxford Review of Education*, *47*(2), 189–206.

Harris, R., & Graham, S. (2019). Engaging with curriculum reform: Insights from English history teachers' willingness to support curriculum change. *Journal of Curriculum Studies*, *51*(1), 43–61.

Hibel, J., & Penn, D. M. (2020). Bad apples or bad orchards? An organizational analysis of educator cheating on standardized accountability tests. *Sociology of Education*, *93*(4), 331–352.

Hinkle, D. E., Wiersma, W., & Jurs, S. G. (2003). *Applied statistics for the behavioral sciences* (4th ed.). Houghton Mifflin.

Hofflinger, A., & von Hippel, P. T. (2020). Missing children: How Chilean schools evaded accountability by having low-performing students miss high-stakes tests. *Educational Assessment, Evaluation and Accountability*, *32*(2), 127–152.

Holloway, J., & Brass, J. (2018). Making accountable teachers: The terrors and pleasures of performativity. *Journal of Education Policy*, *33*(3), 361–382.

Houtsonen, J., Czaplicka, M., Lindblad, S., Sohlberg, P., & Sugrue, C. (2010). Welfare state restructuring in education and its national refractions: Finnish, Irish and Swedish teachers' perceptions of current changes. *Current Sociology*, *58*(4), 597–622.

Jacob, B. A., & Levitt, S. D. (2003). Rotten apples: An investigatinon of the prevalence and predictors of teacher cheating, *Quarterly Journal of Economics*, *118*(3), 843–877.

Jones, M. G., Jones, B. D., & Hargrove, T. Y. (2003). *The unintended consequences of high-stakes testing*. Rowman & Littlefield.

Koretz, D. (2017). *The testing charade*. University of Chicago Press.

Kuhfeld, W. F., Tobias, R. D., & Garratt, M. (1994). Efficient experimental design with marketing research applications. *Journal of Marketing Research*, *31*(4), 545–557.

Lang, A. (1996). Standpoint: The logic of using inferential statistics with experimental data from nonprobability samples: Inspired by Cooper, Dupagne, Potter and Sparks. *Journal of Broadcasting and Electronic Media*, *40*(3), 422–430.

Latham, J. P., & Locke, E. A. (2007). New developments in and directions for goal-setting research. *European Psychologist*, *12*(4), 290–300.

Leeper, T. (2018). *Online experimental methods*. Draft prepared for the Oxford Handbook of Electoral Persuasion.https://s3.us-east-2.amazonaws.com/tjl-sharing/assets/OnlineExperimentalMethods.pdf

Lennert, da, Silva, A. L., & Mølstad, C. E. (2020). Teacher autonomy and teacher agency: A comparative study in Brazilian and Norwegian lower secondary education. *The Curriculum Journal*, *31*(1), 115–131.

Levatino, A. (2021). Surveying Principals and Teachers: Methodological Insights into the Design of the REFORMED Questionnaires. REFORMED Methodological Papers No 2. Doi: 10.5281/zenodo.4450774

Lindquist, E. F. (1951). Preliminary considerations in objective test construction. In E. F. Lindquist (Ed.), *Educational measurement* (2nd ed., pp. 119–158). American Council on Education.

Lingard, B., Martino, W., & Rezai-Rashti, G. (2013). Testing regimes, accountabilities and education policy: Commensurate global and national developments. *Journal of Education Policy*, *28*(5), 539–556.

Locke, E. A. (1968). Toward a theory of task motivation and incentives. *Organizational Behavior and Human Performance*, *3*, 157–189.

Maaranen, K., & Wågsås Afdal, H. (2020). Exploring teachers' professional space using the cases of Finland, Norway and the US. *Scandinavian Journal of Educational Research*, *66*(1), 134–149.

Madaus, G. F., & Clarke, M. (2001). The adverse impact of high-stakes testing on minority students: Evidence from one hundred years of test data. In G. Orfield & M. L. Kornhaber (Eds.), *Raising standards or raising barriers? Inequality and high-stakes testing in public education*. (pp. 85–106) Century Foundation Press.

Maroy, C., & Pons, X. (2019). *Accountability policies in education. A comparative and multilevel analysis in France and Quebec*. Springer.

Maroy, C., & Voisin, A. (2017). *Think piece on accountability: Background paper prepared for the 2017/8 Global Education Monitoring Report* [Research Report]. UNESCO. https://halshs.archives-ouvertes.fr/halshs-01705982/document

Martinelli, C., Parker, S. W., Pérez-Gea, A. C., & Rodrigo, R. (2018). Cheating and incentives: Learning from a policy experiment. *American Economic Journal: Economic Policy*, *10*(1), 298–325.

Mitchell, M. S., Baer, M. D., Ambrose, M. L., Folger, R., & Palmer, N. F. (2018). Cheating under pressure: A self-protection model of workplace cheating behavior. *Journal of Applied Psychology*, *103*(1), 54–73.

Mizala, A., & Schneider, B. R. (2014). Negotiating education reform: Teacher evaluations and incentives in Chile (1990–2010). *Governance*, *27*(1), 87–109.

Mullinix, K. J., Leeper, T., Druckman, J., & Freese, J. (2015). The generalizability of survey experiments. *Journal of Experimental Political Science*, *2*(2), 109–138.

Murnane, R., & Cohen, D. (1986). Merit pay and the evaluation problem: Why most merit pay plans fail and a few survive. *Harvard Educational Review*, *56*(1), 1–18.

Nichols, S. L., & Berliner, D. C. (2005). *The inevitable corruption of indicators and educators through high-stakes testing* (No. EPSL0503-101-EPRU). Education Policy Studies Laboratory, Arizona State University. Retrieved August 9, 2021, from https://files.eric.ed.gov/fulltext/ED508483.pdf

Ohemeng, F., & McCall-Thomas, E. (2013). Performance management and "undesirable" organizational behaviour: Standardized testing in Ontario schools. *Canadian Public Administration*, *56*(3), 456–477.

Osborn, M. (2006). Changing the context of teachers' work and professional development: A European perspective. *International Journal of Educational Research*, *45*(4–5), 242–253.

Patrick, B., Plagens, G. K., Rollins, A., & Evans, E. (2018). The ethical implications of altering public sector accountability models: The case of the Atlanta cheating scandal. *Public Performance & Management Review*, *41*(3), 544–571.

Pino, M. P., Oyarzún, G., & Salinas, I. (2016). A critique to the standardization for accountability: Narrative of resistance of the assessment system in Chile. *Cuadernos Cedes*, *36*, 337–354.

Saldaña, J. (2021). *The coding manual for qualitative researchers*. SAGE.

Schreier, M. (2018). Sampling and generalization. In U. Flick (Ed.), *The SAGE handbook of qualitative data collection* (pp. 84–98). SAGE.

Steiner, P. M., Atzmüller, C., & Su, D. (2016). Designing valid and reliable vignette experiments for survey research: A case study on the fair gender income gap. *Journal of Methods and Measurement in the Social Sciences*, *7*(2), 52–94.

Sullivan, A., Johnson, B., Simons, M., & Tippett, N. (2020). When performativity meets agency: How early career teachers struggle to reconcile competing agendas to become 'quality' teachers. *Teachers and Teaching*, *27*(5), 388–403.

Thiel, C., & Bellmann, J. (2017). Rethinking side effects of accountability in education: Insights from a multiple methods study in four German school systems. *Education Policy Analysis Archives*, *25*(93), 1–32.

Toledo Figueroa, D., & Wittemberg, D. (2015). *Re-shaping teacher careers in Chile: Selected International evidence*. OECD.

Torres, R. (2021). *Does test-based school accountability have an impact on student achievement and equity in education? A panel approach using PISA* (Working Paper No. 250). OECD Education. Retrieved June 23, 2022, from, https://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=EDU/WKP(2021)7&docLanguage=En

UNESCO. (2012). *International standard classification of education ISCED 2011*. United Nations Educational, Scientific and Cultural Organization.

Van Thiel, S., & Leeuw, F. L. (2002). The performance paradox in the public sector. *Public Performance & Management Review*, *25*(3), 267–281.

Verger, A., & Parcerisa, L. (2017). A difficult relationship: Accountability policies and teachers. International evidence and key premises for future research. In: Akiba, M. & LeTendre, G. (eds) International Handbook of Teacher Quality and Policy (pp.241–254). New York: Routledge.

Webb, P. T. (2006). The choreography of accountability. *Journal of Education Policy*, *21*(2), 201–214.

Welsh, D. T., & Ordóñez, L. (2014). The dark side of consecutive high performance goals: Linking goal-setting, depletion and unethical behavior. *Organizational Behavior and Human Decision Processes*, *123*(2), 79–89.

Wengraf, T. (2001). *Qualitative research interviewing: Biographic narrative and semi-structured methods*. SAGE.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin*, *1*(6), 80–83.

Zientek, L., Ebran Yetkiner, Z., Özel, S., & Allen, J. M. (2012). Reporting confidence intervals and effect sizes: Collecting the evidence. *Career and Technical Education Research*, *37*(3), 277–295.

Author Biographies

Antonina Levatino is a researcher in the Department of Sociology of the Universitat Autònoma de Barcelona. Her main research interests concern global trends in education and global governance, focusing, among others, on the enactment and effects of accountability and autonomy reforms in education.

Lluís Parcerisa is an Assistant Professor in the Department of Teaching and Learning and Educational Organization at the Universitat de Barcelona. His main research interests include global governance, policy sociology, the datafication of schooling, and the enactment and effects of school autonomy with accountability reforms in the education sector.

Antoni Verger is Professor of Sociology at the Universitat Autònoma de Barcelona and research fellow at the Catalan Institution for Research and Advanced Studies (ICREA). His research examines educational reform through the perspective of comparative and global policy studies. Over the years, has specialized in the study of educational privatization, school autonomy and accountability reforms.