Educational Assessment, Evaluation and Accountability (2024) 36:107–137
https://doi.org/10.1007/s11092-023-09424-3

Teachers gaming the system: exploring opportunistic behaviours in a low-stakes accountability system

Gerard Ferrer-Esteban[1] · **Marcel Pagès**[2,3]

Received: 29 November 2022 / Accepted: 14 November 2023 / Published online: 6 December 2023

Abstract

Based on the theoretical approaches of social capital and institutional trust, this paper seeks to identify contextual factors and conditions behind teacher behaviours which aim to alter the results of standardised tests in the Italian low-stakes accountability system. Numerous studies report significant factors associated with student cheating, but research into the factors of teacher-led opportunistic actions is scarce. Logistic regression models with fixed-effects at classroom level, with interaction terms, were carried out to identify factors increasing the likelihood of teacher misbehaviour. Models included approximately 79,100 primary, lower and upper secondary classrooms. Indicators of teacher cheating were estimated through algorithms based on suspicious answer strings from standardised tests. The results suggest that teacher cheating may be understood as a form of support for the most vulnerable students, since it is, to a greater extent, found helping low-income students, grade-retained students, as well as students in socially homogenous school settings. The findings also reveal that teacher cheating is consistently related to collectively share non-civic-minded behaviours and practices undertaken by teachers, which do not match legal requirements, such as within-school social segregation and exclusion of students from tests. Heterogeneous effects show that, even in classrooms with external controllers, the lower the civic capital in a school, the more misbehaviour are found. Relevant implications for research, social theory and policy are discussed.

Keywords Low-stakes accountability · Testing · Social capital · Teacher behaviour · Cheating

✉ Gerard Ferrer-Esteban
gferrere@uoc.edu

Marcel Pagès
marcel.pages@uab.cat

[1] Psychology and Educational Sciences Studies, Universitat Oberta de Catalunya, Barcelona, Catalunya, Spain

[2] Department of Pedagogy, Universitat de Girona, Girona, Catalunya, Spain

[3] Department of Sociology, Universitat Autònoma de Barcelona, Bellaterra (Cerdanyola del Vallès), Catalunya, Spain

1 Introduction: accountability, opportunistic behaviours and motivations to game the system

Contemporary education systems are increasingly adopting external evaluation policy instruments for accountability purposes (Verger et al., 2019). Globally, school governance reforms are being adopted differently in attempts to encompass school autonomy, external forms of accountability and administrative control (Ingersoll & Collins, 2017; Verger et al., 2019). Performance-based accountability (PBA) has become one of the main accountability mechanisms that aims to improve the quality of education by making school actors more responsible for the performance of student results in external standardised tests. Teachers and principals are expected to make use of the data derived from the test to reflexively identify aspects of improvement and implement instructional changes to enhance learning and performance. In "higher stakes" systems, although some data use for improvement may be expected, the main accountability rationale is driven by the consequences associated with the test, often attached to a given scheme of incentives and/or sanctions (Maroy, 2015).

Beyond the policy design of accountability systems, as well as the theory of change expected from different models, accountability mechanisms may generate unexpected effects as they alter the perceptions, expectations, and behaviours of school actors (Maroy & Pons, 2019). Instruments have "a life of their own" and tend to gain autonomy from their initial designs, often evolving towards unexpected outcomes (Lascoumes & Le Galès, 2007; Le Galès, 2016). In the case of PBA, unexpected results have been reported in different dimensions of teaching and instruction, including a wide range of 'opportunistic behaviours', understood as responses derived from "perverse incentives" (Ryan, 2003) with negative consequences that push school actors to adopt instrumental actions to improve the performance in standardised tests. Such opportunistic and instrumental behaviours are expected to be more prone in high-stakes accountability regimes, where "teachers are subjected to higher levels of external pressure to achieve better educational outcomes, especially because of the threat of sanctions these systems involve" (Verger & Parcerisa, 2017, p. 246). The most well-known opportunistic behaviours are those practices that are aimed at prioritising student performance over learning, including not only the so-called teaching to the test and narrowing the curriculum (Au, 2007; Berliner, 2011; Ohemeng & McCall-Thomas, 2013) but also forms of direct or indirect cheating (Amrein-Beardsley et al., 2010; Hibel & Penn, 2020; Jacob & Levitt, 2003).[1]

There are many forms of teacher cheating: copying, suggesting correct answers, 'adjustments' while checking answer sheets, etc. Other strategies, not focused on the test, can be altering the composition of students who are being tested: students may be strategically classified as students with special needs (so excluded from aggregated scores) or simply advised not to go to class on the day of the test. There have been several cases of teachers cheating identified in many countries, published mostly in reports and newspaper articles (Amrein & Berliner, 2002; Nichols

[1] For a more detailed review on the effects of accountability systems in education, including undesired and opportunistic behaviours, see Verger and Parcerisa (2017).

& Berliner, 2005), triggering a broad public debate on the reliability of the high-stakes testing system. For example, in 2013, a report sent by the U.S. Government Accountability Office to the Secretary of the U.S. Department of Education described incidents in Illinois, Maryland, Pennsylvania, Texas, Washington D.C. and in California, where the results of 23 schools were invalidated for cheating by school administrators and teachers in 2012. In Chicago schools, it has been estimated that, every year, there is a minimum of 4–5% teacher cheating in elementary schools and that this phenomenon was associated with minimal changes in incentive schemes, which led to significant distortions in conduct (Jacob & Levitt, 2003). In other countries analysed, such as Hungary, similar levels of teacher cheating to that in American schools have been reported (Horn, 2012).

The rationale and motivations behind turning to these opportunistic behaviours depend on the goals of those who cheat, which are conditioned by the characteristics of the accountability and standardised assessment systems (Stecher, 2002). Opportunistic behaviours are expected to be more likely in high-stakes accountability systems, which offer extrinsic incentives encouraging teachers and schools to raise student scores (Jacob & Levitt, 2003). However, explicit consequences cannot be assumed as the single or the main explanatory factor explaining instrumental practices. Lowering the stakes does not necessarily prevent opportunistic behaviours, which have also been reported even in the absence of schemes of explicit incentives and sanctions. Evidence of opportunistic behaviours in systems that do not attach relevant consequences to the test have been found, for instance, in Italy, where significant cheating on tests has been uncovered (Bertoni et al., 2013; Paccagnella & Sestito, 2014; Quintano et al., 2009); or in some German states, where practices of narrowing the curriculum and teaching to the test have been reported (Jäger et al., 2012). According to the sociology of numbers approach, the performative effect of testing and measurement is a sufficient condition for exerting external pressure for school actors, modulating their attitudes, understandings and behaviours (Gorur, 2015; Hardy, 2015). Teachers and principals may also feel pressure for the external test because of the reputational effect of PBA (Camphuijsen, 2021).

To date, there have been few studies aimed at examining the rationale and motivations to cheat in compulsory education (primary and lower education), and even fewer dealings specifically with factors behind cheating led by teachers. Studies have been mainly focused on the relationship between the incentive systems (rewards and sanctions) and teacher behaviours. In high-stakes systems, teachers normally attribute misbehaviour to the pressures of their social environment to get better results (parents and media), but particularly from the educational authorities and the accountability system based on explicit threats of dismissal if they fail. However, what are the factors explaining cheating practices in contexts where test results are not published in schools' league tables, or where there are no formal and direct schemes of incentives and sanctions associated with the average results of teachers and schools?

This study contributes to answering this question by developing a comprehensive framework to interpret the phenomenon of teachers cheating in a low-stakes accountability system. Specifically, it focuses on identifying contextual factors and

conditions behind teacher behaviours aimed at gaming the system in an accountability framework that, on the face of it, offers no motivation to do so.

2 Social capital as collective civic capital to understand cheating practices

Teacher cheating might be understood as an unexpected and undesired professional response. Over the last few decades, an emerging corpus of literature has focused on the social dimension of teaching — understood as a set of professional practices and routines shaped on shared norms, values and beliefs — and embedded in social contexts of professional and personal relations and interactions. Accordingly, a growing body of research has adopted the analytical perspective of "social capital" to understand teachers' practices. However, this literature tends to use the term "social capital" ambiguously, since it is often conceived as an umbrella concept that includes both individual and collective conceptions, contributing to blurring the concept and making it more difficult to use it empirically (Coppe et al., 2022). To overcome such limitations, we explicitly approach social capital as a collectively shared civic capital, based on institutional trust and social reciprocity.

Moreover, such conceptualisation appears to be a very well-suited analytical perspective to better understand the rationale for cheating in low-stakes standardised tests. This approach understands social capital as collective civic engagement based on reciprocity and trust (Gittell & Vidal, 1998; Putnam, 1995, 2000; Woolcock & Narayan, 2000). Citizens in a highly civic-minded community have a high civic engagement, are politically equals and are more prone to act on the basis of solidarity, trust and tolerance, while giving a strong boost to the associations of public life (Putnam, 2000). Here, social capital is understood as a collective stock and refers to the moral obligations and norms, social values (such as trust) and networks (such as voluntary associations) that enable people to act collectively (Woolcock & Narayan, 2000) in favour of the collective benefit (Portes, 2000). This last feature of collective stock (social capital as a feature of communities) and collective benefits means this conception diverges from other conceptions, according to which ties are established to yield benefits to individuals (Coleman, 1988; Portes, 2000). According to Putnam, the collective accumulation of social capital drives political integration and economic development to higher levels (Putnam, 1995). Here, social capital is understood as a civic culture that can be collectively used and it has a strong cultural base (Trigilia, 2011).

This approach to social capital can be distinguished into two different forms. First, *bonding social capital*, which is established between members of a community with a homogeneous composition; second, *bridging social capital*, which is set among different social groups or socially heterogeneous groups. In both forms of social capital, the networks and the associated norms of reciprocity are valuable in terms of trust, solidarity and mutual support. Both society-based and community-based networks may provide resources to support the most disadvantaged members or groups.

However, to be considered as a stock of capital, any form of capital should have a positive economic payoff, should be measurable and should have defined mechanisms through which social capital can be accumulated and depreciated (Solow, 1995). As it is potentially exclusive, since socially homogeneous groups have self-referenced interests which may diverge from those held by other groups, or from those of society as a whole, the bonding form of social capital may have negative consequences. In these cases, this form of social capital may be detrimental to the *bridging* form, which is concerned with solidarity, mutual respect and cooperation — values related to the welfare of the society. Indeed, as pointed out by Portes (1998), the strong ties, which bring benefits to members of a group, generally restrict access to outsiders. Therefore, the potential negative externalities of the bonding form of social capital cast doubt on this approach's suitability.

To meet Solow's criteria, Guiso et al. (2011) propose another definition of social capital as civic capital: "those persistent and shared beliefs and values that help a group overcome the *free rider* problem in the pursuit of socially valuable activities" (p. 419). The authors refer to a culture-based civic capital where values and beliefs are shared by a community and persist over time. This civic capital is related to all types of economic interactions and not restricted to political participation (Guiso et al., 2010). The authors point out that relevant direct measures of civic capital may identify values that induce people to be against actions that give private benefits at high social costs. Specifically, they refer to opinions about *free riding* and other behaviours which deviate from the public good (e.g. labour absenteeism, tax evasion or avoidance and littering).

3 Objectives and research questions

The main objective of this paper is to explore the factors that contribute to explain why teacher cheating in standardised tests is observed in low-stakes accountability systems. Along with the above-mentioned approach of social capital, we aim to find the factors associated with teacher cheating, which are distributed in three aggregated levels of analysis: classroom, school and province (Table 1). We can divide our research questions and hypotheses into two sub-groups, according to different explanatory factors we aim to investigate.

First, we want to observe to what extent cheating behaviours are more likely in classrooms with higher stocks of bonding social capital (see Table 1, columns a and b). We also want to confirm whether teacher cheating, understood as a form of community-based support, is addressed to help more disadvantaged students (e.g. low-SES, socioeconomic status or grade-retained students). Strong ties established in socially homogeneous groups may benefit their members, which, in this case, is in the form of teacher help for students who have difficulties with the test, while being detrimental to society at large — since teacher cheating undermines the monitoring and accountability objectives of the testing systems.

Second, we want to observe to what extent teachers in contexts with lower levels of civic-minded capital are more likely to cheat. We use indicators of behaviours collectively shared that deviate from the public good (of the whole society); that is, actions

Table 1. Teacher cheating-related factors and operationalisation of variables

Aggregation of variables	Variables	Teacher cheating-related factors		
		Bonding social capital		Lack of civic social capital
		Conditions for cheating	Rationale for cheating (compensation function)	Contextual factors associated with cheating practices
		[a]	[b]	[c]
Classroom level	Homogeneity of social composition	✓		
	Classrooms with low SES average		✓	
	Classrooms with grade-retained students		✓	
School level	Exclusion of students from tests			✓
	Social tracking between classrooms			✓
Province level	Teacher absenteeism			✓

Source: authors.

that provide private benefits at high social cost: with regard to school-level behaviours and practices, we analyse to what extent cheating behaviours are more likely in schools which undertake practices that do not match legal requirements or recommendations, such as social tracking of students and exclusion of students from tests (Table 1, column c). As regards the context-level behaviours, we explore whether teachers in schools located in provinces with a lack of civic capital — higher rates of teacher absenteeism — are more prone to cheat in standardised tests. Then, even if deterrents such as external controllers during the test taking have been proved to be effective in preventing cheating, we also analyse whether such deterrents are effective in schools where non-civic-minded values are deeply rooted.

4 Data and method

4.1 The Invalsi student performance dataset

Italy is a good setting for exploring the factors associated with opportunistic behaviours in low-stakes systems, due to the richness of the data coming from the standardised testing system. In Italy, teacher cheating has already been associated with the geographical location of schools (Quintano et al., 2009), certain locally shared values (Paccagnella & Sestito, 2014), student peer effects (Lucifora & Tonello, 2012) or deterrent mechanisms (Bertoni et al., 2013), such as the external control during tests.

We used the dataset of standardised tests administered by the Italian National Institute for the Evaluation of the Education System (Istituto nazionale per la valutazione del sistema educativo di istruzione e di formazione, Invalsi), from the 2011/12 academic year. This study dealt with surveys of 5th, 6th and 10th grade student performance. In the Italian education system structure, these grades correspond, respectively, to the 5th year of primary education, the 1st year of lower secondary education and the 2nd year of upper secondary. The national survey was obligatory for all schools in 2009 and 2010. For this research, in the 2011/12 wave, the estimation models constructed covered approximately 1,426,000 students and 79,100 classrooms from the 5th, 6th and 10th grades (on average, 10, 11 and 15 years old), spread over five macro-areas, 20 regions and 103 provinces. The tests, which were not high-stakes, covered mathematics and Italian language and were administered by teachers following a protocol set by Invalsi. This protocol suggested that the presence of teachers not specialised in the subject being tested would be appropriate. External inspectors were sent to a sample of classrooms, in schools randomly selected across the regions, to control the realisation of tests, to check the answer sheets and return the results to Invalsi.

4.2 Empirical analysis: determinants, incentives and deterrents of cheating

We estimated logistic regression models to explore to what extent the explanatory factors included in the equation make the likelihood of cheating increase or decrease. The baseline equation can be expressed as follows:

$$\begin{aligned} logit\left(cheat_{csp}\right) =& \alpha + \gamma_1 O_{csp} + \gamma_2 H_{csp} \\ &+ \gamma_3 F_{csp} + \gamma_4 E_{sp} + \gamma_5 T_{sp} \\ &+ \gamma_6 A_p + \gamma_7 D_p + \gamma_8 C_{csp} \\ &+ \gamma_9 P_p + \gamma_{10}\left(O_{csp} * T_{sp}\right) \end{aligned} \quad (1)$$

where $cheat_{csp}$ is whether a classroom c in school s in province p is suspected of cheating. O_{csp} covers two dummies that refer to the opportunity to cheat, that is, whether a classroom c was monitored by an external controller and whether it was indirectly controlled. The rest of the variables included in the model, whose descriptive statistics are shown in Table 2, are described below.

A cheating indicator based on suspicious answer strings The binary dependent variable $cheat_{csp}$ is whether a classroom is suspected of cheating in mathematics and reading tests. In the literature, research reports and articles propose or compare different methods to identify student cheating in multiple choice tests (Angoff, 1974; Belleza & Belleza, 1989; Frary, 1993; Sotaridona & van der Linden, 2006; Wesolowsky, 2000). In contrast, there are very few methods to identify teacher cheating. The most relevant is that elaborated by Jacob and Levitt (2003), which is the method that we partially replicated. This method to detect opportunistic behaviours adopted by teachers combines two indicators. The first is *Unexpected Test Score Fluctuations*, which basically refers to unexpected score gains that can be explained by cheating. A classroom will be suspected of cheating if unexpectedly large gains are followed by lower than usual test scores for the same students the following year. When test scores are monitored over time, student gains due to talented teachers or rich educational programs are likely to be permanent. This first indicator could not be calculated due to the lack of longitudinal data.

Instead, we focused on the second: a composite indicator of Suspicious Answer Strings.[2] With this indicator, different ways for a teacher to cheat could be detected — not only the easiest but also more sophisticated actions. The Suspicious Answer Strings indicator is a combination of four measures: an unlikely block of identical answers given to consecutive questions in the classroom, the classroom average variance across all test items, the variance (as opposed to the mean) in the degree of correlation across questions within a classroom and the extent to which a student's response pattern differs from other students with the same aggregate score that year. The overall measure of cheating is constructed, within a given subject and grade, ranking classrooms on each of the four indicators and taking the sum of squared ranks across the four measures (see the appendix for more information about the construction of the cheating indicator based on suspicious answer strings). In our

[2] Refer to the appendix for further details on the construction of the cheating indicator derived from suspicious answer strings.

Table 2. Descriptive statistics

	Primary education					Lower secondary education					Upper secondary education				
	Mean	*SD*	Min.	Max.	Freq.	*Mean*	*SD*	Min.	Max.	Freq.	*Mean*	*SD*	Min.	Max.	Freq.
Classroom suspected of cheating (90th percentile cut-off)	0.11	0.31	0	1		0.11	0.31	0	1		0.11	0.31	0	1	
Unmonitored classrooms in unmonitored schools	0.76	0.42	0	1	19,453	0.70	0.46	0	1	16,567	0.57	0.49	0	1	11,827
Monitored classrooms	0.07	0.25	0	1	1676	0.08	0.27	0	1	1925	0.10	0.30	0	1	2132
Unmonitored classrooms in monitored schools	0.17	0.37	0	1	4305	0.22	0.41	0	1	5124	0.32	0.47	0	1	6702
Social student heterogeneity in classroom	0.86	0.26	0.1	2		0.88	0.24	0.1	2		0.86	0.16	0.3	2	
Low-SES classrooms	0.33	0.47	0	1	8478	0.33	0.47	0	1	7872	0.33	0.47	0	1	6921
Mid-SES classrooms	0.33	0.47	0	1	8478	0.33	0.47	0	1	7872	0.33	0.47	0	1	6864
High-SES classrooms	0.33	0.47	0	1	8478	0.33	0.47	0	1	7872	0.33	0.47	0	1	6876
Fraction of grade-retained students in classroom	0.03	0.05	0	1		0.07	0.08	0	1		0.20	0.18	0	1	
Students who missed the test	10.92	5.12	0	48		11.37	5.24	0	52		14.53	8.48	0	60	
Social tracking between classrooms	0.28	0.15	0	2		0.27	0.14	0	1		0.28	0.11	0	1	
Fraction of first-generation immigrants in classroom	0.05	0.07	0	1		0.06	0.07	0	1		0.05	0.07	0	1	
Fraction of second-generation immigrants in classroom	0.05	0.09	0	1		0.05	0.08	0	1		0.07	0.08	0	1	
Fraction of female students in classroom	0.50	0.11	0	1		0.49	0.11	0	1		0.50	0.29	0	1	
Classroom size	17.76	3.91	10	32		19.99	3.97	10.0	34		19.12	4.56	10.0	35	
School size	5.11	2.08	1	14		6.13	3.06	1	17		7.97	3.27	1	20	
Teacher absenteeism (province)	7.22	1.63	4.1	13		7.28	1.67	4.1	13		7.29	1.67	4.1	13	
Unemployment rate	8.44	3.99	2.1	19		8.54	4.01	2.1	19		8.65	4.00	2.1	19	
Province with metropolitan area	0.39	0.49	0	1		0.36	0.48	0.0	1		0.36	0.48	0.0	1	
Index of teacher precariousness	-0.20	1.09	−4.9	1		−0.22	1.10	−4.9	1		−0.27	1.14	−4.9	1	
Share of early school leavers	0.20	0.07	0.1	0		0.20	0.07	0.1	0		0.20	0.07	0.1	0	
Share of teacher turnover	0.10	0.02	0.1	0		0.10	0.02	0.1	0		0.10	0.02	0.1	0	

Source: own elaboration, based on data from the Invalsi dataset 2011–12.

empirical analysis, we employ the 90th percentile cut-offs to pinpoint classrooms suspected to cheating, with the 95th percentile used in robustness checks:[3]

Social capital-related factors as predictors of cheating In our study, we explored factors associated with civic social capital. The measure H_{csp} represents classroom social homogeneity, calculated based on the standard deviation of the SES index — socioeconomic status — within each classroom. The SES index was calculated and provided by the Italian National Institute for the Evaluation of the Education System, using data about students' parents' education, occupation and household possessions. The standard deviation as a measure of social homogeneity offers insights into the socioeconomic diversity of students in a classroom: a lower standard deviation would suggest more homogeneity (similar backgrounds), while a higher standard deviation would indicate more heterogeneity (diverse backgrounds).

F_{csp} is a vector capturing factors indicative of the compensation function of teacher cheating, which includes the classroom proportion of students who were retained a grade and a dummy for whether a classroom has a low SES average. Classroom with a low SES were identified by generating quartiles from the classroom-level SES average. Classrooms in the bottom quartile were assigned a value of 1, while the others were assigned a value of 0.

We have also included three school-related factors associated to the social capital of teachers and other school agents: the school-level measure E_{sp}, which represents the fraction of students who missed the test (as a proxy of the percentage of students excluded from the test taking); the school-level measure T_{sp}, which is a variable denoting the extent of school practices that result in social tracking between classrooms and the province-level measure A_p, which is the share of teacher absenteeism. Here, we were also interested in analysing the extent to which the mentioned external deterrents of cheating during the test taking are effective in schools with high levels of non-civic-minded culture or lack of civic capital. For this we included the vector $O_{csp} \times T_{sp}$, which represents the interaction term between a continuous measure of a lack of civic capital at the school level (social tracking between classrooms within schools) and the binary predictors of both direct and indirect external control during the test.

4.3 Addressing omitted-variable bias

We deal with potential problems of omitted variables, since behaviours denoting law acceptance or conformity (e.g. tax compliance vs. tax evasion) may be effectively driven by factors not necessarily related to civic capital, such as economic payoffs or legal enforcement. This is the reason why, as indirect measures of civic capital, outcome-based measures such as behaviours, are difficult to interpret (Guiso et al., 2010). Although we assume that measures of deviant behaviours may

[3] Findings from the robustness checks can be provided upon request.

only partially inform the lack of civic-minded capital, we argue that, to the extent that the following two assumptions are accepted, they may still be important predictors. Firstly, when dealing with school- and province-level aggregated factors, such as social tracking of students or tax evasion, we assume that the legal framework and the measures of law enforcement are held constant. Since we exploit within-country variability of a single country, we partially account for the driving force of legal deterrents: measures of legal enforcement, recommendations, deontological professional standards, etc. Secondly, we assume that much of the motivation and incentive to act opportunistically, associated with the social structure and territorial characteristics, can be captured through the socioeconomic composition controls, as well as by the territorial fixed effects (regions and macro-areas).

When talking about social dissimilarity between classrooms, we should consider that, in Italy, especially in urban centres, many schools occupy several buildings which are often physically separated — possibly by more than 1 km. This may lead to social homogeneity within, and social heterogeneity between, classrooms, not as a result of practices of student tracking, but as a reflection of the different social composition of the surroundings of the school buildings. Bearing in mind that no information is available on the urban or rural location of schools, we account for this phenomenon, albeit partially, by including measure Z_{sp}:school size, which reflects the number of classrooms in each school.

The variables used as controls measures in the equations were at classroom, school and province level. At the classroom level, we included the vector C_{csp}, which covers aggregated student background data: the fraction of female students and the fraction of first-generation and second-generation immigrant students in the classroom. We also controlled for the classroom size S_{csp}, expressed as the number of students enrolled. At the province level, we accounted for a range of social, economic and geographic characteristics, which was included in the vector P_p of province characteristics data: GDP per capita, unemployment rate, whether provinces contained a metropolitan area, population size and density of province and share of adult population participating in education. In this vector, we also included province-aggregated data related to the school system, such as the share of early school leavers, an index of teacher precariousness (teachers with a temporary contract) and the share of teacher turnover. Finally, in order to control for systemic and cultural differences at the territorial level, we separately introduced a set of macro-area and region fixed-effects. We do not add an explicit error term as the logit transformation inherently accounts for an error structure. Standard errors were clustered, in separate specifications, at the province and the school levels.

4.4 Results reporting and goodness of fit

The results are interpreted in two different ways, accounting for the non-linear nature of the logit analysis: through an *odds ratio* and by using *average marginal effects*. The *odds ratio* indicates the odds of a classroom being suspected of cheating ($y = 1$) relative to the odds of it not being suspected of cheating (y

= 0). Specifically, if p is the probability of $y = 1$, then the odds are given by the ratio $p/(1 - p)$. The *average marginal effects* were calculated estimating the average of the classroom marginal effects (expressed as a percentage), indicating how an increase in x is associated with an increase or decrease of the probability of y being equal to 1 (classroom suspected of cheating). While for dummy variables, the marginal effect is expressed in comparison to the base; for continuous variables, it is expressed for one-unit change in the explanatory factor. The marginal effects on the probability of being suspected of cheating, calculated as the average of the classroom marginal effects, are given by:

$$\frac{\partial p}{\partial x_j} = \frac{\sum F'\left(x'\beta\right)}{n}\beta_j \tag{2}$$

The goodness of fit was measured with the percentage of values correctly predicted. First of all, taking the estimated coefficient $\hat{\beta}$, we calculate the predicted probability $\hat{p}$ that y would be equal to 1 for each classroom (i.e. that cheating had occurred) in the dataset:

$$\hat{p} = pr\left[y = 1|x\right] = F\left(x'\hat{\beta}\right) \tag{3}$$

The predicted probabilities $\hat{p}$ in logit and probit models are limited between 0 and 1 and indicate the likelihood that $y = 1$. Once we have $\hat{p}$ to check the good fit of the model, we calculate the percentage of values correctly predicted. This is the proportion of true predictions to total predictions ($\hat{y} = y$). Our model is a good fit if we correctly give at least 70% true or correct predictions. The models derived from Eq. (3) correctly predict between 88% (90th percentile cut-off) and 94% of values (95th percentile cut-off), and therefore we can confirm that they are suitable to use for the empirical analysis.

5 Main results

In this section, we test our hypotheses and present the main results according to the social capital analytical approach to understand opportunistic behaviours.

5.1 Cheating and the bonding form of social capital

As mentioned earlier, if a context is socially more homogeneous, stronger ties between teachers and students are more plausible. The first indicator is a measure of social dispersion within classrooms (classroom social heterogeneity), which, inverted, becomes a proxy of strong ties established in a close and socially homogeneous context. Since we refer specifically to teacher cheating, we do not approach classroom social homogeneity to identify interactions among students as a determinant of cheating (Lucifora & Tonello, 2012), but rather in contexts where ties of

mutual support between teacher and students are more likely to be found. As can be seen in Table 3, the social dispersion indicator is a robust predictor of teacher cheating, net of classroom, school and territorial controls. Specifically, looking at the marginal effects, we observed that a one-unit increase in the standard deviation of the students' socioeconomic index within classrooms decreases the probability of a teacher being suspected of cheating. For a one-unit increase in social heterogeneity, we expect the probability of cheating to be reduced by 1.6% in primary and 3% in lower and upper secondary. In terms of the odds ratio, if we invert the scale, we can say that for a one-unit decrease in the measure of social dispersion, the probability for being suspected of cheating is 1.2 and 1.4 times more likely than not being suspected, in primary and in lower and upper secondary, respectively. Overall, teachers are more prone to support students, suggest answers or fill in the answer sheets when the classroom composition is socially more homogeneous, where bonding forms of social capital are more likely.

Context of teacher cheating: a compensation function for disadvantaged students From the bonding approach of social capital, cheating can be understood as a strategy targeted to support socially and academically disadvantaged students. If we look at the fraction of grade-retained students in classrooms, we observe that teacher cheating is significantly associated with these academically disadvantaged students in the upper secondary for mathematics and Italian and in the lower secondary only for mathematics. Indeed, a one-percent unit increase of grade-retained students in upper secondary school increases the probability of a teacher cheating by 12% in Italian (Table 3) and 8% in mathematics.[4] In the case of lower secondary teachers, the marginal effect increases to 8% for a one-percent unit increase in retained students. This is, however, an outcome that may be influenced by the differential presence of retained students depending on the education grades. The extent to which we move up through the grades, the percentage of grade-retained students increases. It is almost non-existent in primary, representing only 3.4% of students. In the lower secondary this is about 7%, while in the upper secondary, it rises to almost 20% (Table 2).

In addition, to test whether cheating is more likely in classrooms with socially disadvantaged students, we estimated the probability that being suspected of cheating was dependent on the socioeconomic average status of classrooms. Here, we compared classrooms with a high SES average with low- and mid-SES classrooms. The results are significant and robust in all specifications: teachers in classrooms with low social composition are more prone to be suspected of cheating. Specifically, in primary and upper secondary, having a low-SES composition doubles the probability of cheating in both Italian and mathematics (Table 3). In terms of average marginal effects, teachers in low-SES classrooms are 5.5 to 7% more likely to cheat than teachers in high-SES classrooms. These results are attenuated, though still statistically significant, in the 1st year of lower secondary school and between mid-SES and high-SES classrooms.

[4] Tables presenting the results on mathematics are available upon request.

Table 3. Teacher cheating and bonding social capital-related factors (Italian language). Interactions with external control

Classroom suspected of cheating	Coefficients			Odds-ratio			Average marginal effects			Interaction terms		
	Primary	Lower sec.	Upper sec.	Primary	Lower sec.	Upper sec.	Primary	Lower sec.	Upper sec.	Primary	Lower sec.	Upper sec.
Social student heterogeneity in classroom	−0.190* (0.10)	−0.333*** (0.11)	−0.351** (0.14)	0.827* (0.08)	0.717*** (0.08)	0.704** (0.10)	−0.016** (0.01)	−0.030*** (0.01)	−0.031** (0.01)	−0.178* (0.10)	−0.326*** (0.11)	−0.349** (0.14)
Mid-SES classroom (ref: low-SES classrooms)	−0.381*** (0.06)	−0.107* (0.06)	−0.201*** (0.07)	0.683*** (0.04)	0.898* (0.05)	0.818*** (0.05)	−0.032*** (0.00)	−0.010* (0.01)	−0.018*** (0.01)	−0.378*** (0.06)	−0.107* (0.06)	−0.202*** (0.07)
High-SES classroom (ref: low-SES classrooms)	−0.656*** (0.08)	−0.134** (0.06)	−0.813*** (0.10)	0.519*** (0.04)	0.874** (0.06)	0.444*** (0.04)	−0.055*** (0.01)	−0.012** (0.01)	−0.072*** (0.01)	−0.654*** (0.08)	−0.136** (0.06)	−0.812*** (0.10)
Fraction of grade-retained students in classroom	0.664 (0.48)	0.389 (0.30)	1.317*** (0.17)	1.943 (0.93)	1.475 (0.45)	3.733*** (0.62)	0.056 (0.04)	0.035 (0.03)	0.117*** (0.01)	0.653 (0.48)	0.386 (0.30)	1.317*** (0.17)

Table 3. (continued)

Classroom suspected of cheating	Coefficients			Odds-ratio			Average marginal effects			Interaction terms		
	Primary	Lower sec.	Upper sec.	Primary	Lower sec.	Upper sec.	Primary	Lower sec.	Upper sec.	Primary	Lower sec.	Upper sec.
Students who missed the test	0.005 (0.01)	0.013*** (0.00)	0.018*** (0.01)	1.005 (0.01)	1.013*** (0.00)	1.018*** (0.01)	0.000 (0.00)	0.001*** (0.00)	0.002*** (0.00)	0.006 (0.01)	0.013*** (0.00)	0.018*** (0.01)
Social tracking between class-rooms	0.329* (0.17)	0.541*** (0.18)	0.465** (0.20)	1.390* (0.24)	1.718*** (0.30)	1.591** (0.31)	0.028* (0.01)	0.049*** (0.16)	0.041** (0.02)	0.258 (0.19)	0.392** (0.20)	0.622** (0.28)
Controllers in class-room during the test taking (ref: no control)	−1.194*** (0.15)	−0.636*** (0.09)	−0.998*** (0.13)	0.303*** (0.05)	0.530*** (0.05)	0.369*** (0.05)	−0.100*** (0. 01)	−0.058*** (0. 01)	−0.089*** (0.01)	−1.751*** (0.27)	−0.998*** (0.21)	−0.972*** (0.26)
Controllers in school during the test taking (ref: no control)	−0.328*** (0.08)	−0.154** (0.07)	−0.275*** (0.07)	0.720*** (0.06)	0.858** (0.06)	0.760*** (0.05)	−0.027*** (0. 01)	−0.014** (0. 01)	−0.024*** (0.01)	−0.332*** (0.16)	−0.283** (0.11)	−0.119 (0.20)

Table 3. (continued)

Classroom suspected of cheating	Coefficients			Odds-ratio			Average marginal effects			Interaction terms		
	Primary	Lower sec.	Upper sec.	Primary	Lower sec.	Upper sec.	Primary	Lower sec.	Upper sec.	Primary	Lower sec.	Upper sec.
Social tracking between classrooms × controllers in classroom during the test taking										1.608** (0.69)	1.125** (0.55)	−0.086 (0.86)
Social tracking between classrooms × controllers in school during the test taking										0.017 (0.41)	0.411 (0.34)	−0.119 (0.20)
Observations (classrooms)	23,518	22,619	19,534	23,518	22,619	19,534	23,518	22,619	19,534	23,518	22,619	19,534
Pseudo R^2	0.1664	0.1010	0.1314	0.1664	0.1010	0.1314				0.1667	0.1012	0.1315

Table 3. (continued)

Source: own elaboration, based on data from the Invalsi dataset 2011–12. Robust standard errors clustered by school in parentheses. The dependent variable is the cheating indicator using the 90th percentile cut-off. The control variables include classroom, school and province factors. Classroom and school: fraction of female students, the fraction of first-generation and second-generation immigrant students in the classroom, classroom size and school size. Province: GDP per capita, unemployment rate, whether provinces contained a metropolitan area, population size and density of province, share of adult population participating in education, share of adult population with a low level of education, share of early school leavers, an index of precarious teachers, the share of teacher turnover and the share of teacher absenteeism. It includes macro-area fixed effects. Region fixed effects were included in separate specifications, but the sign and significance of the results were not affected (results available on request).

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

5.2 Cheating as lack of civic capital

In this section, we tested the hypothesis that teachers in contexts with lower levels of civic capital are more prone to cheat. As mentioned earlier, here, we specifically explore to what extent cheating behaviours are associated with collectively-shared misbehaviours, which are detrimental to the public good. Beside deontological and ethical considerations, we consider them as non-civic-minded behaviours since they do not comply with legal requirements or are against laws and regulations.

School-level practice: excluding students from the test-taking In Italy, students can take an adapted test or be exempt from testing, if they have a certified disability or a specific learning disability certification, such as dyslexia or dyscalculia. While this information is not available, we have got a proxy, albeit an imperfect one, for students who have been opportunistically excluded from the test to avoid worse results or to inflate the average results. This proxy is obtained by the difference between the number of students taking the test and the number of students on the class record. While the number of students with certified disabilities may vary significantly from class to class, or from school to school, this information should be independent from the teacher's likelihood of cheating. Thus, any association observed between the increase in students absent in the test-taking and the likelihood of a teacher being suspected of cheating would point out simultaneity of opportunistic strategies to modify the average test results.

In Table 3, we see how the number of students absent from taking the tests is associated with the probability of cheating by both lower and upper secondary school teachers. With primary education the only exception, teachers that directly alter the test results to improve test scores are also prone to exclude a higher proportion of students from test taking as an instrumental strategy. This association shows how different strategic behaviours teachers use to inflate scores can run in parallel, even in low-stakes systems. This result is in line with those obtained by studies that have reported opportunistic behaviours related to the exclusion of low-SES and special-education-needs students, including cases such as The Netherlands (Mons, 2009), Canada (Bélair, 2005, cited by Mons, 2009), the USA (Cullen & Reback, 2006; Figlio & Getzler, 2006; Haladyna et al., 1991; Jacob, 2005; Madaus et al., 1992) and Chile (Hofflinger & von Hippel, 2020).

School-level practice: within-school social segregation The second factor associated with teacher cheating is the practice of non-random allocation of students across classrooms based on their social background. In short, schools that undertake tracking practices, which result in social polarisation of classrooms, are also more likely to show higher levels of teacher cheating. A one-unit increase in the standard deviation of a classroom's socioeconomic index within schools increases the probability of the classroom being suspected of cheating by between 3 and 5% in Italian (Table 3), and between 2 and 5% in mathematics. As for the odds-ratio, for a one-unit increase in the measure of social tracking, the probability of being suspected of cheating is, on average across grades, 1.6 times more likely than not being suspected. It is worth noting that in the case of upper secondary education, no significant association was found in mathematics. This is probably because upper secondary schools are already highly

tracked, having a more homogeneous social composition. On the other hand, the most solid association between cheating and social tracking, as a proxy of non-minded-civic behaviour, is in the first year of the lower secondary school: this is when students from primary school are allocated into classrooms by the school board.

Civic capital and deterrents to cheat At this point, we questioned to what extent deterrents of cheating are effective when considering the level of non-civic-minded engagement of schools. Firstly, it is worth noting that external control of the classroom during the test is the most important deterrent of cheating (Table 3). Moreover, we also confirm the spill-over effects of the presence of external controllers in other classrooms in the same school. Now, we introduce interaction terms to test the change in the interaction effects of those measures of external control on teacher cheating, depending upon the degree of civic capital in schools. Figures 1 and 2 show that the strong association between the lack of civic capital and teacher cheating is also observed in classrooms

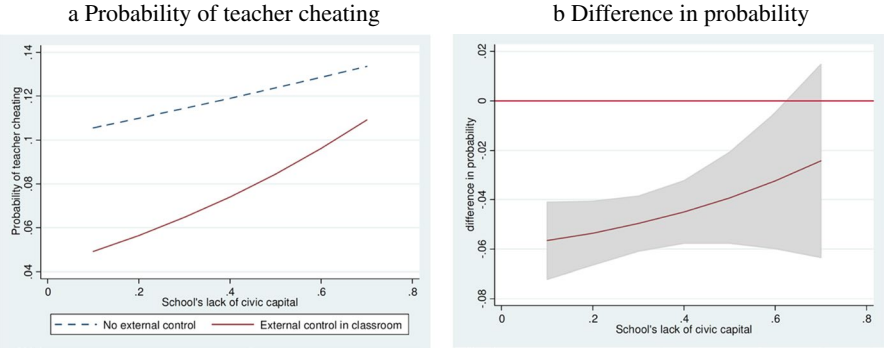

Fig. 1. Probability of teacher cheating: interaction between lack of civic capital in schools and external control in classrooms (direct control). Source: own elaboration, based on data from the Invalsi dataset 2011-12

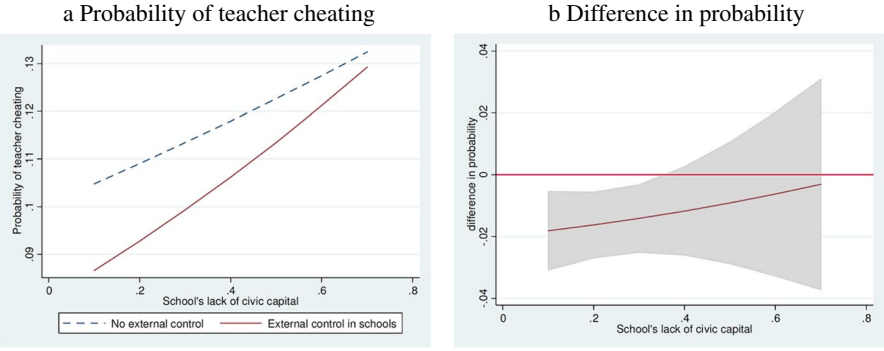

Fig. 2. Probability of teacher cheating: interaction between lack of civic capital in schools and external control in schools (indirect control). Source: own elaboration, based on data from the Invalsi dataset 2011-12

with direct and indirect external control. Although the predicted probability is always higher in unmonitored classrooms, even in classrooms with external control, the less civic capital in a school, the more cheating behaviours are found. If we compare the association trends between monitored and unmonitored classrooms, we observe that they tend to converge when schools show high levels of non-civic-minded behaviours. This means that, especially in the case of indirectly controlled classrooms, the predicted probability of cheating is the same as in unmonitored classrooms.

In Fig. 1, the difference in probability between classrooms with and without direct external control is not statistically significant when the levels of civic capital are very low. This means that direct control in classrooms during test taking is an effective deterrent, except in those schools with very low levels of civic capital. Conversely, Fig. 2 indicates that indirect control is an effective deterrent when the school has medium or high levels of civic-minded capital. Specifically, the difference between the levels of teacher cheating between unmonitored and indirectly monitored classrooms is no longer significant when the lack of civic capital is in the 80th percentile (0.37).

Context-related factor: teacher absenteeism in the province The rate of teacher absenteeism may be approached not only as a school system-related factor but also as a context-related factor. In either case, it is a powerful predictor of a lack of civic capital, as the results indicate a solid association between teacher cheating and the extent to which teachers are absent from the school, significant and robust across the grades and in both tests. Regarding the magnitude of the association, the tables show that, when the rate of teacher absenteeism at the province level increases by a one-percent unit, the probability of a teacher being suspected of cheating increases between 1.3 and 1.8%, depending on both the grade and the test (Table 4). Even when controlling for regional dummies (instead of macro-area), the association is significant, with the sole exception of the mathematics test in the upper secondary school.

Here, we confirm that the context in which schools are located is strongly correlated with cheating: teachers are more likely to cheat in contexts where more deviant behaviours — behaviours against laws and regulations — are collectively shared. This remains significant even after accounting for a wide range of province characteristics, including the socioeconomic composition, and for cultural and economic differences across macro-areas.

6 Discussion

This paper contributes to the educational policy debate by providing a comprehensive framework for school administrators and policymakers to understand and explore potential explanatory factors behind teacher cheating and other opportunistic behaviours. We have empirically studied the factors behind opportunistic behaviours in education systems with schemes of low-stakes accountability. More specifically, we have explored the practices of teacher cheating through the analytical lens of

Table 4. Teacher cheating and lack of civic capital: teacher absenteeism (Italian language)

Classroom suspected of cheating	Teacher absenteeism					
	Primary education	Lower secondary	Upper secondary	Primary education	Lower secondary	Upper secondary
Teacher absenteeism (province)	0.217***	0.150***	0.161***	0.184***	0.078***	0.130***
	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.04)
Social student heterogeneity in classroom	−0.190*	−0.333***	−0.351**	−0.173*	−0.311***	−0.339**
	(0.10)	(0.11)	(0.14)	(0.10)	(0.11)	(0.14)
Mid-SES classroom (ref: Low-SES classrooms)	−0.381***	−0.107*	−0.201***	−0.371***	−0.093	−0.165**
	(0.06)	(0.06)	(0.07)	(0.06)	(0.06)	(0.06)
High-SES classroom (ref: Low-SES classrooms)	−0.656***	−0.134**	−0.813***	−0.643***	−0.130*	−0.766***
	(0.08)	(0.06)	(0.10)	(0.08)	(0.07)	(0.10)
Share of grade-retained students in classroom	0.664	0.389	1.317***	0.573	0.278	1.400***
	(0.48)	(0.30)	(0.17)	(0.49)	(0.31)	(0.16)
Students who missed the test	0.005	0.013***	0.018***	0.004	0.013***	0.019***
	(0.01)	(0.00)	(0.01)	(0.01)	(0.00)	(0.00)
Social tracking between classrooms	0.329*	0.541***	0.465**	0.298*	0.517***	0.520**
	(0.17)	(0.18)	(0.20)	(0.17)	(0.18)	(0.20)
Controllers in classroom during the test taking (ref: no control)	−1.194***	−0.636***	−0.998***	−1.100***	−0.550***	−0.884***
	(0.15)	(0.09)	(0.13)	(0.14)	(0.09)	(0.13)
Controllers in school during the test taking (ref: no control)	−0.328***	−0.154**	−0.275***	−0.226***	−0.065	−0.175***
	(0.08)	(0.07)	(0.07)	(0.07)	(0.07)	(0.06)
Macro-area fixed effect	Yes	Yes	Yes	No	No	No
Region fixed effect	No	No	No	Yes	Yes	Yes
Observations (classrooms)	23,518	22,619	19,534	23,518	22,619	19,534
Pseudo R^2	0.1664	0.1010	0.1314	0.1741	0.1061	0.1365

Source: own elaboration, based on data from the Invalsi dataset 2011–12. Robust standard errors clustered by province in parentheses. The dependent variable is the cheating indicator using the 90th percentile cut-off. The control variables include classroom, school and province factors. Classroom and school: classroom SES average, fraction of female students, the fraction of first-generation and second-generation immigrant students in the classroom, classroom size and school size. Province: GDP per capita, unemployment rate, whether provinces contained a metropolitan area, population size and density of province, share of adult population participating in education, share of adult population with a low level of education, share of early school leavers, an index of precarious teachers and the share of teacher turnover.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

civic social capital. Our results challenge the assumption that the formal incentives and sanctions of high stakes accountability systems are the single driver of opportunistic behaviours. We show, contrary to traditional thinking within the field of educational accountability and assessment, how in low-stakes educational systems, teacher cheating can be associated with factors related to social capital. Our results suggest that cheating and free-rider behaviours can be explained as both stock and lack of civic social capital.

First, we have explored whether cheating behaviours are more likely in contexts where forms of bonding social capital prevail. The results of our analysis confirm that cheating is more prone to happen in social homogeneous environments: where strong ties are more likely, cheating levels appear to be higher and oriented to help the most socially and academically disadvantaged students — students with a low socioeconomic status and the grade-retained — in what could be understood as an altruistic attitude. Our findings are concurrent with existing research in the Italian context. Studies have suggested that the higher levels of cheating in Southern Italy are due to both a lower endowment of bridging social capital and a higher degree of bonding social capital (Bertoni et al., 2013). Other studies have demonstrated that cheating is positively associated with measures of particularistic values (Paccagnella & Sestito, 2014), which can be also understood as civic values related to bonding social capital. For instance, Paccagnella and Sestito (2014) found strong associations between cheating and contexts where people use close local networks to a greater extent.

In parallel, we have explored how and to what extent teachers' cheating behaviours are associated with both school and context-based behaviours. We have hypothesised that teachers in schools and communities with lower levels of civic capital are more prone to cheat. Here, teacher cheating is understood as a misbehaviour that would reflect collective non-civic-minded beliefs and values. This has been highlighted by Paccagnella and Sestito (2014), who showed how cheating is negatively associated with proxies of trust towards education authorities and non-adherence to the rule of law. In our paper, findings confirm how teachers in schools with lower levels of civic capital (school-level non-civic-minded practices) or within contexts which have a lack of civic capital (territorially-aggregated levels of teacher absenteeism) are significantly more likely to cheat.

Another relevant and original contribution of this paper refers to the significant and robust correlation between teacher cheating and school-based opportunistic practices. Specifically, our results suggest that teachers are more likely to cheat in schools that also undertake practices ignoring institutional recommendations from authorities or legal requirements, such as student tracking or the exclusion of students from tests. These results suggest that teachers are more prone to cheat in institutional contexts where other opportunistic behaviours disregarding formal regulations and prescriptions are also found.

The practice of excluding students from tests has been already identified internationally, especially in countries with high-stakes tests such as Chile, the USA and the Netherlands. In the Netherlands, for example, the inspectorate has reported how, in some cases, students who were more likely to be sent to less prestigious school tracks did not take part in the high-stakes tests (Mons, 2009). In the case of

Ontario, Canada, teachers reported that some schools had consistently been reducing the number of students to bring up the average (Bélair, 2005, cited by Mons, 2009). In Chile, schools serving disadvantaged students inflated their scores by having low-performing students miss high-stakes tests (Hofflinger & von Hippel, 2020). In the USA, numerous cases have been identified in several states such as California, Florida, Illinois, Texas and Alabama, where low-ability students were excluded to raise average scores (Haladyna et al., 1991; Madaus et al., 1992). For instance, several authors found that teachers responded strategically to incentives by classifying certain students having special education needs (Cullen & Reback, 2006; Figlio & Getzler, 2006; Jacob, 2005). Our results make a contribution to this literature, adding that strategic behaviours teachers use to inflate scores can run in parallel, even in low-stakes systems, like that of Italy. More research is needed to elucidate whether these practices are also correlated with other forms of opportunistic behaviours in different settings and how it is related to diverging stocks of civic social capital.

The results of our study are also relevant in terms of identifying potential deterrents of teacher cheating. Our results show that low levels of civic engagement and trust are associated with opportunistic behaviours and cheating practices. While we have found that direct control in classrooms is an effective deterrent of teacher cheating, for those schools with very low aggregated levels of social capital, this was not the case. Similarly, schools with lower levels of civic social capital are also the exception in terms of external control as a hindering factor for cheating. In other words, mechanisms that are expected to be effective in limiting cheating are essentially unproductive in these low social civic capital environments. Therefore, more efforts are needed to develop institutional trust and civic capital rather than on developing external control systems without sufficient support for teachers and schools.

7 Implications and conclusions

Altogether, the results presented suggest several implications for research, social theory and policy. First, regarding the implications for research, the evidence provided shows how the undesired effects of accountability systems are not only associated with the formal consequences attached to the test. In this regard, the existing literature tends to highlight the distinction between high and low-stakes accountability systems. This dichotomous differentiation might be flawed for different reasons. The consequences defined in the test might be formally clear but appear to be more blurred in the actual realities of classrooms and schools. Accountability system designs typically consider formal and material consequences, although informal, symbolic and reputational impacts tend to emerge when the test is implemented in real-school settings. For this reason, more than a dichotomous approach, we should understand accountability stakes as a continuum, considering both the formally designed stakes and the unexpected impacts and consequences emerging during the process of implementation.

We have shown how opportunistic behaviours and teacher cheating are not only a singular effect of high stakes accountability regimes. These results reinforce the idea of the performative effect of testing, numbers and metrics as powerful tools to

change and modulate the perceptions and behaviours of school actors, who either elude their control or adapt and respond to diverse accountability demands. Such considerations invite us to further investigate under what circumstances, contexts and policy designs different accountability instruments generate particular effects in classrooms and schools. Future research should address the relationships between policy design, context and effects, analysing how similar accountability designs generate variegated effects in different contexts, as well as the other way around: how different designs might generate similar results depending on the context where they are implemented. Comparative, qualitative and quasi-experimental designs may be a promising research strategy to better understand the complex nature of such phenomena.

Our paper also suggests important contributions for social theory. The findings reinforce the interactionist approach according to which deviance may be understood as a socially constructed process linked to rules of behaviour created in particular social sub-groups, where illicit behaviours are interiorised and justified to respond to exogenously determined situations. Civic capital is also linked to moral standards as defined by the social cognitive theory. People avoid behaving in a detrimental way when they self-regulate and activate internal controls in accordance with such standards. Moral standards, which are socially modelled, are translated into actions and behaviours through which moral agency is exercised (Bandura et al., 1996). However, a self-regulatory system is not invariant, so conduct may differ significantly even if moral standards remain constant. The agreement between moral standards and effective behaviours may support the use of misbehaviours as proxies for non-civic-minded values with consequences on the public good.

A particular paradox between universalistic and particularistic understandings of social justice and civic behaviours is also identified in our research. Such tension is at the very core of the problem analysed and is both relevant for policy and theory. According to Bandura et al. (1996), detrimental conduct may be considered personally and socially acceptable depending on the extent to which they are portrayed in the service of valued social or moral purposes. This is a key factor in the process of moral justification of such behaviour and could be reinforced by an adverse socioeconomic context, exogenously determined and perceived as unfair. In this sense, cheating may simultaneously generate not only particular benefits but also general damage: some opportunistic actions may be undertaken for the benefit of the most disadvantaged, generating not only a particular benefit for those in more vulnerable conditions but also harming the implementation of accountability policies, possibly even including their eventual equity purposes from a universal understanding of social justice.

To limit and balance such tension, accountability systems might need to be revisited towards more *intelligent* designs (Crooks, 2003; Ehren et al., 2020; Lingard, 2009; O'Neill, 2013). In this sense, more flexibility is needed to ensure fairer designs and protect the general purpose of the accountability system — i.e. inform to improve and enhance more quality and equity for education systems — and, at the same time, ensure a proper systematisation to protect its internal and external validity. To advance towards such a horizon, accountability schemes should be oriented towards processes of quality improvement rather than external

tools of control, involving different actors, but specially ensuring the role of experts and professionals, to enhance a system based on relations of trust among different actors. Where mistrust and perceptions of unfairness will be in place, policymakers should assume that opportunistic behaviours and cheating will prevail, limiting the capacity of accountability systems to properly work for quality improvement and the enhancement of educational justice.

Appendix 1. Construction of the cheating indicator based on suspicious answer strings.

Measure 1: unlikely block of identical answers on consecutive questions

For the first measure, Jacob and Levitt (2003) predicted the likelihood that each student would provide a specific answer to each question based on past test scores, future test scores and background characteristics. Since we did not have access to longitudinal data, we relied solely on background information — including immigration background, SES status and gender. We employed a multinomial logit for each item on the exam (where only one option among multiple choices was correct) to predict student responses. This model was estimated using data from other students in the same grade and subject:

$$\Pr\left(Y_{isc}=j\right)=\frac{e^{\beta_j x_s}}{\sum_{j=1}^{J} e^{\beta_j x_s}} \tag{4}$$

where Y_{isc} represents the response of student s in class c on item i. The possible responses (J) range from 3 to 5, and x_s is a vector including the socio-demographic variables for student s. The predicted probability of a student choosing a specific response is determined by the likelihood of other students (in the same grade and subject) with similar background characteristics choosing that same response. We are unable to account for future and prior test scores, which unfortunately increase the likelihood of identifying exceptional teachers as cheaters. Instead, we estimate the probability of selecting each potential response, rather than estimating the likelihood of selecting the correct one, given relevant background variables. As Jacob and Levitt (2003) highlight, we take advantage of any additional information that is provided by particular response patterns within a classroom. Then, using the estimates derived from this model, we determined the predicted probability that each student would answer item i in the way they did, proving one measure per student s for each item i:

$$p_{isc}=\frac{e^{\hat{\beta}_k x_s}}{\sum_{j=1}^{J} e^{\hat{\beta}_j x_s}} \quad \text{for } k=\text{response actually chosen by student } s \text{ on item } i. \tag{5}$$

By taking the product of probabilities across items for each student, we calculated the probability that a student would consecutively answer a string of questions from item m to item n as observed:

$$p_{sc}^{mn} = \prod_{i=m}^{n} p_{isc} \tag{6}$$

We took the product over all students in the classroom c, who had identical responses in the string. Then the product is:

$$\tilde{p}_{sc}^{mn} = \prod_{s \in \left\{ z : S_{ic}^{mn} = \overline{S}_{sc}^{mn} \right\}} p_{sc}^{mn} \tag{7}$$

where z is defined as a student; s_{sc}^{mn} as the string of responses for student z from item m to item n; and $\overline{s}_{sc}^{mn}$ as the string for student s. $\tilde{p}_{sc}^{mn}$ collapses to p_{sc}^{mn} for each student, and there will be ns distinct values within the class, to the extent that there are ns students in classroom c, and each student has a unique set of responses to these particular items. On the contrary, if students in class c have identical answers, only one value of $\tilde{p}_{sc}^{mn}$ will be found. This calculation has been repeated for all possible strings of length three to seven. Finally, the indicator of the least likely block of identical answers given on consecutive questions is created taking the minimum of the predicted block probability for each classroom:

$$measure\ 1_c = \min_s \left(\tilde{p}_{sc}^{mn} \right)$$

Measure 2: classroom average variance across all test items

The second measure is designed to capture general patterns of similarity in student responses. Its construction involves several steps. First, residuals are calculated for each of the possible choices that a student could make for each item:

$$e_{jisc} = 0 - \frac{e^{\hat{\beta}_j x_s}}{\sum_{j=1}^{J} e^{\hat{\beta}_j x_s}} \text{if } j \neq k$$

$$1 - \frac{e^{\hat{\beta}_j x_s}}{\sum_{j=1}^{J} e^{\hat{\beta}_j x_s}} \text{if } j = k \tag{8}$$

where e_{jisc} is the residual for response j on item i by student s in classroom c.

Then, information for each student is combined to create a classroom level measure of the response to item i. First, we sum the residuals for each response across students within a classroom. This value should be near to zero if there is no within-class correlation:

$$e_{jic} = \sum_{s} e_{jisc} \tag{9}$$

Second, we sum across the possible responses for each item within classrooms. Then, we square each of the residual measures to accentuate outliers, and divide by the number of students in the classroom (ns_c) to normalize by class size:

$$v_{ic} = \frac{\sum_j e_{jic}^2}{ns_c} \tag{10}$$

The statistic v_{ic} captures the variance of student responses to item i within classroom c. To highlight the tendencies in response patterns at the classroom level, we first aggregate the residuals for each response across students, and then sum the classroom level measures for each response, rather than initially summing across responses within each student. The second measure of suspicious strings is determined by taking the classroom average (across items) of this variance term across all test items:

$$measure\ 2_c = \bar{v}_c = \frac{\sum_i v_{ic}}{ni}$$

where ni is the number of items on the exam.

Measure 3: variance in the degree of correlation across questions within a classroom

$$measure\ 3_c = \sigma_{v_c}^2 = \frac{\sum_i \left(v_{ic} - \bar{v}_c\right)^2}{ni}$$

Measure 4: variability in student response patterns

The fourth indicator assesses the degree to which an individual student's response pattern deviates from others who achieved the same aggregate score that year. Let q_{isc} denote whether student s in classroom c answered item i correctly (1 if correct, and 0 otherwise). Let A_s represent the aggregate score of student s for the exam. The objective is to determine the proportion of students at each aggregate score level who answered each item correctly. If ns_A denotes the number of students with an aggregate score of A, then this proportion, represented as $\overline{q}_i^A$, can be expressed as:

$$\overline{q}_i^A = \frac{\sum_{s \in \{z : A_z = A_s\}} q_{isc}}{ns_A} \tag{11}$$

Then, it is calculated a measure to quantify the deviation of student s response pattern from that of other students who achieved the same aggregate score. To do this, the student's answer on item i is subtracted from the mean response of

all students with an aggregate score of A. These deviations are then squared and summed across all items on the exam:

$$Z_{sc} = \sum_i \left(q_{isc} - \overline{q}_i^A \right)^2 \tag{12}$$

The final indicator is calculated by subtracting the mean deviation for all students with the same aggregate score, $\overline{Z}^A$ and then summing the results for the students within each classroom:

$$measure\ 4_c = \sum\nolimits_s \left(Z_{sc} - \overline{Z}^A \right)$$

As mentioned previously, the overall measure of cheating is constructed, for a given subject and grade, by ranking classrooms based on each of the four indicators. The overall score is then determined by taking the sum of the squared ranks across these four measures. In our empirical analysis, we employ the 90th percentile cut-offs to pinpoint classrooms suspected to cheating, with the 95th percentile used in robustness checks[5]:

$$cheating_{cdg} = \left(\text{rank\_}measure1_{csg}\right)^2 + \left(\text{rank\_}measure2_{csg}\right)^2 + \left(\text{rank\_}measure3_{csg}\right)^2 + \left(\text{rank\_}measure4_{csg}\right)^2$$

where $cheating_{cdg}$ indicates whether a classroom c in subject d in grade g is suspected of cheating.

Acknowledgements This study was initially conducted as part of a project on the Italian school evaluation system, promoted by the Fondazione Agnelli (Italy), and was further developed within the framework of the European Research Council-funded REFORMED project. Gerard Ferrer-Esteban would like to especially thank Martino Bernardi (Fondazione Agnelli) and Gianfranco De Simone (Evaluation Lab of the Fondo per la Repubblica Digitale) for helpful suggestions on econometric modelling, and the rest of the Agnelli Foundation's staff for helpful general comments and discussions. He is grateful to Stefano Molina (Unione Industriali di Torino) and Jorge Rodríguez Menés (Universitat Pompeu Fabra) for providing helpful theoretical insights to further understand cheating behaviour. He is also grateful to participants in workshops at Irvapp (Trento) and Invalsi (Rome) for comments, to Patrizia Falzetti (Invalsi) for help with the data, and Sergio Longobardi (Università degli Studi di Napoli 'Parthenope') for providing information on the cheating detection method used by the Invalsi. Marcel Pagès wishes to acknowledge the support received from the 'Margarita Salas' postdoctoral research program.

Funding Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature. This work was initially supported by the Fondazione Agnelli (Italy), and subsequently by the European Research Council under Grant 680172 (REFORMED project).

Data availability The data that support the findings of this study is not publicly available as it is protected by the Italian data protection law. It can be obtained through an application procedure at the Italian National Institute for the Evaluation of the Education System (Istituto nazionale per la valutazione del sistema educativo di istruzione e di formazione, Invalsi). Information on how to obtain it and reproduce the analysis is available from the corresponding author on request.

Declarations

Conflict of interest The authors declare no competing interests.

[5] Findings from the robustness checks can be provided upon request.

References

Amrein, A. L., & Berliner, D. C. (2002). *An analysis of some unintended and negative consequences of high-stakes testing*. EPSL-0211-125-EPRU Education Policy Research Unit, Arizona State University Retrieved November 20, 2022, https://nepc.colorado.edu/sites/default/files/EPSL-0211-125-EPRU.pdf

Amrein-Beardsley, A., Berliner, D. C., & Rideau, S. (2010). Breaking professional law: Degrees of cheating on high stakes tests. *Education Policy Analysis Archives, 18*(14). https://doi.org/10.14507/epaa.v18n14.2010

Angoff, W. H. (1974). The development of statistical indices for detecting cheaters. *Journal of the American Statistical Association, 69*(345), 44–49. https://doi.org/10.2307/2285498

Au, W. (2007). High-stakes testing and curricular control: A qualitative metasynthesis. *Educational researcher, 36*(5), 258–267. https://doi.org/10.3102/0013189X07306523

Bandura, A., Barbaranelli, C., Caprara, G. V., & Pastorelli, C. (1996). Mechanisms of moral disengagement in the exercise of moral agency. *Journal of Personality and Social Psychology, 71*(2). https://doi.org/10.1037/0022-3514.71.2.364

Bélair, L. M. (2005). Les dérives de l'obligation de résultats ou l'art de surfer sans planche. In C. Lessard & P. Meirieu (Eds.), *L'obligation de résultats en éducation. Evolutions, perspectives et enjeux internationaux* (pp. 179–187). De Boeck.

Belleza, F. S., & Belleza, S. F. (1989). Detection of cheating on multiple-choice tests by using error similarity analysis. *Teaching of Psychology, 16*(3), 151–155. https://doi.org/10.1207/s15328023top1603_15

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. https://doi.org/10.1080/0305764X.2011.607151

Bertoni, M., Brunello, G., & Rocco, L. (2013). When the cat is near, the mice won't play: The effect of external examiners in Italian schools. *Journal of Public Economics, 104*, 65–77. https://doi.org/10.1016/j.jpubeco.2013.04.010

Camphuijsen, M. K. (2021). *From trust in the profession to trust in results: A multi-scalar analysis of performance-based accountability in Norwegian education*. (Doctoral dissertation, Universitat Autònoma de Barcelona Retrieved November 20, 2022, https://hdl.handle.net/10803/672512

Coleman, J. S. (1988). Social capital in the creation of human capital. *American Journal of Sociology, 94*, Supplement: Organizations and Institutions: Sociological and Economic Approaches to the Analysis of Social Structure, S95–S120. https://doi.org/10.1086/228943

Coppe, T., Thomas, L., Pantić, N., Froehlich, D. E., Sarazin, M., & Raemdonck, I. (2022). The use of social capital in teacher research: A necessary clarification. *Frontiers in Psychology, 2703*. https://doi.org/10.3389/fpsyg.2022.866571

Crooks, T. (2003). *Some criteria for intelligent accountability applied to accountability in New Zealand*. Paper presented at the Annual meeting of the American Educational Research Association April 22, 2003. Retrieved November 20, 2022, from https://www.fairtest.org/some-criteria-intelligent-accountabi%20lity-applied-a/

Cullen, J. B., & Reback, R. (2006). Tinkering toward accolades: School gaming under a performance accountability system. In *National Bureau of Economic Research* (p. 12286). NBER Working Paper. https://doi.org/10.1016/S0278-0984(06)14001-8

Ehren, M., Paterson, A., & Baxter, J. (2020). Accountability and trust: Two sides of the same coin? *Journal of Educational Change, 21*(1), 183–213. https://doi.org/10.1007/s10833-019-09352-4

Figlio, D. N., & Getzler, L. S. (2006). Accountability, ability and disability: Gaming the system? In T. J. Gronberg & D. W. Jansen (Eds.), *Improving school accountability* (Vol. *14*, pp. 35–49). Emerald Group Publishing. https://doi.org/10.1016/S0278-0984(06)14002-X

Frary, R. B. (1993). Statistical detection of multiple-choice answer copying: Review and commentary. *Applied Measurement in Education, 6*(2), 153–165. https://doi.org/10.1207/s15324818ame0602_4

Gittell, R., & Vidal, A. (1998). *Community organizing. Building social capital as a development strategy*. Sage.

Gorur, R. (2015). Assembling a sociology of numbers. In M. Hamilton, B. Maddox, & C. Addey (Eds.), *Literacy as numbers: Researching the politics and practices of international literary assessment* (pp. 1–16). Cambridge University Press.

Guiso, L., Sapienza, P., & Zingales, L. (2010). Civic capital as the missing link. NBER Working Paper, 15845. *National Bureau of Economic Research*. Retrieved November 20, 2022, from https://www.nber.org/system/files/working_papers/w15845/w15845.pdf

Guiso, L., Sapienza, P., & Zingales, L. (2011). Civic capital as the missing link. *Handbook of social economics, 1*, 417–480. https://doi.org/10.1016/B978-0-444-53187-2.00010-3

Haladyna, T. M., Nolen, S. B., & Haas, N. S. (1991). Raising standardized test scores and the origins of test score pollution. *Educational Researcher, 20*(5), 2–7. https://doi.org/10.3102/0013189X020005002

Hardy, I. (2015). Data, numbers and accountability: The complexity, nature and effects of data use in schools. *British Journal of Educational Studies, 63*(4), 467–486. https://doi.org/10.1080/00071005.2015.1066489

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. https://doi.org/10.1177/0038040720927234

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. https://doi.org/10.1007/s11092-020-09318-8

Horn, D. (2012). Catching cheaters in Hungary estimating the ratio of suspicious classes on the national assessment of basic competencies tests. In *Institute of Economics, Research Centre for Economic and Regional Studies (Hungarian Academy of Sciences)*. Department of Economics (Eötvös Lóránd University).

Ingersoll, R. M., & Collins, G. J. (2017). Accountability and control in American schools. *Journal of Curriculum Studies, 49*(1), 75–95. https://doi.org/10.1080/00220272.2016.1205142

Jacob, B. A. (2005). Accountability, incentives and behavior: The impact of high-stakes testing in the Chicago public schools. *Journal of Public Economics, 89*(5-6), 761–796. https://doi.org/10.1016/j.jpubeco.2004.08.004

Jacob, B. A., & Levitt, S. D. (2003). Rotten apples: An investigation of the prevalence and predictors of teacher cheating. *The Quarterly Journal of Economics, 118*(3), 843–877. https://doi.org/10.1162/00335530360698441

Jäger, D. J., Maag Merki, K., Oerke, B., & Holmeier, M. (2012). Statewide low-stakes tests and a teaching to the test effect? An analysis of teacher survey data from two German states. *Assessment in education: Principles, policy & practice, 19*(4), 451–467. https://doi.org/10.1080/0969594X.2012.677803

Lascoumes, P., & Le Galès, P. (2007). Introduction: Understanding public policy through its instruments—From the nature of instruments to the sociology of public policy instrumentation. *Governance, 20*(1), 1–21. https://doi.org/10.1111/j.1468-0491.2007.00342.x

Le Galès, P. (2016). Performance measurement as a policy instrument. *Policy Studies, 37*(6), 508–520. https://doi.org/10.1080/01442872.2016.1213803

Lingard, B. (2009). Testing times: The need for new intelligent accountabilities for schooling. *QTU Professional Magazine, 24*, 13–19.

Lucifora, C., & Tonello, M. (2012). Students' cheating as a social interaction: Evidence from a randomized experiment in a national evaluation program. In *IZA Discussion Paper, 6967*. Study of Labor (IZA). https://doi.org/10.2139/ssrn.2170655

Madaus, G., West, M., Harmon, M., Lomax, R., & Viator, K. (1992). *The influence of testing on teaching math and science in grades 4-12*. Center for the Study of Testing, Evaluation, and Educational Policy.

Maroy, C. (2015). Comparing accountability policy tools and rationales. In H.-G. Kotthoff & E. Klerides (Eds.), *Governing educational spaces* (pp. 35–56). SensePublishers. https://doi.org/10.1007/978-94-6300-265-3_3

Maroy, C., & Pons, X. (2019). *Accountability policies in education. A comparative and multilevel analysis in France and Quebec*. Springer. https://doi.org/10.1007/978-3-030-01285-4

Mons, N. (2009). *Theoretical and real effects of standardised assessment*, background paper to the study: National testing of pupils in Europe: Objectives, organisation and use of results. EACEA, Eurydice.

Nichols, S. L., & Berliner, D. C. (2005). *The inevitable corruption of indicators and educators through high-stakes testing*. EPSL-0503-101-EPRU Education Policy Research Unit, Arizona State University Retrieved November 20, 2022, https://nepc.colorado.edu/sites/default/files/EPSL-0503-101-EPRU-exec.pdf

O'Neill, O. (2013). Intelligent accountability in education. *Oxford Review of Education, 39*(1), 4–16. https://doi.org/10.1080/03054985.2013.764761

Ohemeng, F., & McCall-Thomas, E. (2013). Performance management and "undesirable" organizational behaviour: Standardized testing in Ontario schools. *Canadian Public Administration, 56*(3), 456–477. https://doi.org/10.1111/capa.12030

Paccagnella, M., & Sestito, P. (2014). School cheating and social capital. *Education Economics, 22*(4), 367–388. https://doi.org/10.1080/09645292.2014.904277

Portes, A. (1998). Social capital: Its origins and applications in modern sociology. *Annual Review of Sociology, 24*, 1–24 https://www.jstor.org/stable/223472

Portes, A. (2000). The two meanings of social capital. *Sociological Forum, 15*(1), 1–12. https://doi.org/10.1023/A:1007537902813

Putnam, R. D. (1995). Bowling alone: America's declining social capital. *Journal of Democracy, 6*(1), 65–78. https://doi.org/10.1353/jod.1995.0002

Putnam, R. D. (2000). *Bowling alone: The collapse and revival of American community*. Simon & Schuster.

Quintano, C., Castellano, R., & Longobardi, S. (2009). A fuzzy clustering approach to improve the accuracy of Italian student data. An experimental procedure to correct the impact of outliers on assessment test scores. *Statistica & Applicazioni*, VII(2), 149-171.

Ryan, J. E. (2003). *The perverse incentives of the No Child Left Behind Act* (pp. 03–17). *UVA School of Law, Public Law Working Paper*. https://doi.org/10.2139/ssrn.476463

Solow, R. (1995). Review of Francis Fukuyama, Trust: The social virtues and the creation of prosperity. *The New Republic, 213*, 37–39.

Sotaridona, L. S., & van der Linden, W. J. (2006). Detecting answer copying when the regular response process follows a known response model. *Journal of Educational and Behavioral Statistics, 31*(3), 283–304. https://doi.org/10.3102/10769986031003283

Stecher, B. M. (2002). Consequences of large-scale, high-stakes testing on school and classroom practice. In L. S. Hamilton, M. B. Stecher, & S. P. Klein (Eds.), *Making sense of test-based accountability in education* (pp. 79–100) Retrieved November 20, 2022, https://www.rand.org/content/dam/rand/pubs/monograph_reports/2002/MR1554.pdf

Trigilia, C. (2011). Capitale sociale tra economia e sociologia: avanti con giudizio. In G. de Blasio & P. Sestito (Eds.), *Il capitale sociale: che cos'è e che cosa spiega* (pp. 29–42). Donzelli.

Verger, A., Fontdevila, C., & Parcerisa, L. (2019). Reforming governance through policy instruments: How and to what extent standards, tests and accountability in education spread worldwide. *Discourse: Studies in the Cultural Politics of Education, 40*(2), 248–270. https://doi.org/10.1080/01596306.2019.1569882

Verger, A., & Parcerisa, L. (2017). A difficult relationship: Accountability policies and teachers. International Evidence and Premises for Future Research. In M. Akiba & G. K. LeTendre (Eds.), *International handbook of teacher quality and policy* (pp. 241–254). Routledge. https://doi.org/10.4324/9781315710068

Wesolowsky, G. O. (2000). Detecting excessive similarity in answers on multiple choice exams. *Journal of Applied Statistics, 27*(7), 909–921. https://doi.org/10.1080/02664760050120588

Woolcock, M., & Narayan, D. (2000). Social capital: Implications for development theory, research, and policy. *The World Bank Research Observer, 15*(2), 225–249. https://doi.org/10.1093/wbro/15.2.225