Wößmann, Ludger

Working Paper

Central Exams Improve Educational Performance: International Evidence

Kieler Diskussionsbeiträge, No. 397

Provided in Cooperation with:
Kiel Institute for the World Economy – Leibniz Center for Research on Global Economic Challenges

Suggested Citation: Wößmann, Ludger (2002) : Central Exams Improve Educational Performance: International Evidence, Kieler Diskussionsbeiträge, No. 397, Institut für Weltwirtschaft (IfW), Kiel

This Version is available at:
https://hdl.handle.net/10419/17922

KIELER DISKUSSIONSBEITRÄGE

397 KIEL DISCUSSION PAPERS

Central Exams Improve Educational Performance: International Evidence

by Ludger Wößmann

CONTENTS

- International comparative studies of student performance have initiated political discussions in countries all over the world on how to improve the educational achievement of students. The empirical evidence suggests that central exams help to achieve higher student performance.
- Central exams direct the incentives of all educational actors towards furthering students' knowledge. By providing the education system with performance information they improve the monitoring of students, teachers, schools, administrators, and parents.
- Using an international micro database of nearly half a million students, this paper finds that students in countries with central exit-exam systems perform substantially better in their middle-school years in both math and science than students in countries without central exams. In quantitative terms, their advantage is 35 to 47 percent of an international standard deviation in test scores, or roughly the equivalent of one year of schooling. The beneficial effect increases as students advance through middle school.
- Good and bad students alike perform better in central-exam systems. In math, the gain of high-performing students is slightly larger than that of low-performing students. There is some evidence that central-exam systems equalize educational opportunities for students from different parental backgrounds.
- School autonomy in budgetary and salary decisions is detrimental in systems without central exams but turns around to be beneficial in systems with central exams. Thus, central exams seem to be a prerequisite for a decentralized system of autonomous schools to achieve high performance.
- The efforts of teachers and students are more concentrated on the goals of the education system when central exams are in place, and parental involvement becomes more informed and effective. Thus, central exams exert their effects through several different impact channels by changing the behavior of different actors in the education process. Given the shortcomings of most school- and teacher-based accountability systems and the substantially higher costs of most resource-based policies, central-exam systems seem a highly attractive policy alternative.

Contents

Paper prepared for the conference "Taking Account of Accountability: Assessing Politics and Policy" at the John F. Kennedy School of Government, Harvard University, June 10-11, 2002. Financial support for the construction of the TIMSS-Repeat micro database by the Program on Education Policy and Governance (PEPG), Harvard University, under a grant from the John M. Olin Foundation is gratefully acknowledged. The hospitality of both the PEPG and the National Bureau of Economic Research during spring 2002, which enabled me to write this paper, is highly appreciated. I would also like to thank Caroline Hoxby and Kathryn Schiller for their constructive discussion of the paper at the conference, John Bishop for his kind provision of data on central exit exams in most of the countries analyzed, Eugenio Gonzalez of the TIMSS International Study Center for helpful clarifications on the TIMSS-Repeat data, Andreas Ammermüller for research assistance in the construction of the TIMSS-Repeat database, and Marty West for many helpful comments on a first draft of this paper.

1 Introduction

Examination systems constitute a vital part of accountability systems in schools. Virtually all education systems examine students' educational achievement—only that this examination takes place in very different ways. A pivotal feature of the execution of exams is whether they are designed, carried out, and graded by individual teachers or whether they are conducted by an entity external to schools. In external-exam systems, every student takes the very same tests, thus making the central exams an intrinsic part of the school system. These exams, which are usually administered by a public agency, tend to be based on the schools' curriculum and grade student performance into multiple levels of achievement based on an external standard, not just relative to students in a class. While often referred to as central exams, "central" need not necessarily mean that the exams are administered by the national government; it can also refer to centralization at some regional level. Being external, neither the teachers nor the students can determine or know the specific questions contained in the exams. To improve performance, it is necessary to teach, or respectively learn, the whole curricular standards on which the exams are based. The external exams may be given in each grade in primary and/or secondary school, in several grades, or—as in the special case considered in the empirical part of this paper—they may take the form of exit exams administered at the end of secondary education, with a minimum score generally required for graduation. The incentives that students, teachers, schools, administrators, and parents face differ substantially between external-exam systems and teacher grading. This paper analyzes these differences and assesses their impact on the functioning of the education system and ultimately on students' academic performance.

The analysis draws on new international evidence from two large cross-country comparative studies of student performance, TIMSS and TIMSS-Repeat.[1] The original TIMSS study was conducted in 1994/95. TIMSS-Repeat was conducted in 1998/99, with the data only recently made available. Data on nationally representative samples of middle-school students are available for 39 countries in TIMSS and for 38 countries in TIMSS-Repeat, with 23 countries participating in both studies. Students were tested in math and science, two central areas in the curriculum of any education system. The data used in this paper includes performance data in both math and science for about 450,000 individual students, as well as background data on families, school resources, and institutional settings for individual students, teachers, and schools. This rich micro database allows the estimation both of how students perform in education systems with and without central exams, and of whether the extent of school autonomy, teacher influence, and parental involvement have different consequences in education systems with and without central exams.

The recent international comparative studies of student performance have initiated political discussions all over the world on how to improve the educational achievement of students. Especially in Germany, the ability of central-exam systems to affect student performance is hotly debated. Some German states (Länder) favor central exams, others do not—currently, seven of the sixteen German states have implemented central exit-exam systems. Divisions on the topic of central exams are apparent both between and within political parties; some political pressure groups favor central-exam systems, others—usually including teacher unions (e.g., GEW Hannover 2002; GEW Hessen 2002)—oppose them. The discussions are often linked to suggestions of greater school autonomy in other areas of decision-making. Also, the question is often raised whether the impact of central exams may differ for students from different family backgrounds. However, these discussions, and educational

[1]The original meaning of TIMSS was "Third International Mathematics and Science Study," as it followed two individual mathematics studies and two individual science studies that had been conducted between 1964 and 1984. TIMSS has since been renamed the "Trends in International Mathematics and Science Study," as assessments are now meant to be conducted on a regular basis every four years. The TIMSS-Repeat study is also known under the acronyms TIMSS-R and TIMSS 1999.

policies in general, are mostly based on preexisting convictions, under general neglect of facts and empirical evidence. A thorough analysis of the cross-country evidence can inform these policy debates.

1.1 Information, Monitoring, Incentives, and Behavior

Accountability systems generally consist of three components: performance standards, measurement of student performance, and consequences for measured performance. Central-exam systems are a specific way of measuring performance, usually against some predefined standards, that do not necessarily have to have explicit consequences attached to the tests. In contrast to many accountability systems currently discussed in the United States that set explicit monetary rewards or sanctions in response to performance, such as school-based accountability systems with monetary consequences for schools or merit-pay systems with monetary consequences for teachers, central-exam systems usually do not set monetary rewards or sanctions themselves. Instead, they rely on the "spontaneous" behavior of the different stakeholders in the education process, thereby working mostly indirectly through implicit consequences.[2]

Most importantly, central exams provide *information* on how individual students perform relative to the national (or regional) student population. This information is not given in the absence of central exams, when classroom teachers grade their students. In the latter setting, performance is generally not comparable across classrooms, and nobody knows whether a mark earned in one class reflects the same scope of contents as a mark earned in another class. In contrast, the information provided by central-exam systems signals the performance of students, teachers, and schools, and it thus facilitates the monitoring of the behavior of the different stakeholders in the education system. Given that the whole education process is fraught with agency problems where principals cannot directly observe what their agents are doing—giving agents leeway to act "opportunistically"—this role performed by central exams may be pivotal to how the education system works. The information they create may be used by a lot of stakeholders in education—and even beyond the actual education system. Rewards for capable and striving teachers may come from their heads of school who now are able to monitor teachers' performance, and rewards for studious students may come from the labor market, where potential employers or institutions of higher education now have the necessary information to compare different students' performance. Likewise, lazy students may be penalized in the labor market, and both teachers and schools may be pressured by parents and administrators, who now possess the necessary information to evaluate their performance. In effect, central exams thoroughly change the incentives faced by the different stakeholders in education, focusing incentives on student learning.

In terms of teaching and learning, a pivotal difference in the incentive mechanism of central exams relative to teacher-set exams is that neither teachers nor students know beforehand which specific questions are going to be asked. Teachers therefore cannot "get away" with skipping whole content areas in the classroom. They are instead forced to teach the whole subject areas as prescribed in the standards and cannot effectively scale down the standards. Furthermore, if well implemented, the possibility of teacher cheating—for example by discussing the specific questions of the exam beforehand or by telling students that certain content areas will not be covered in the exam—is eliminated.

It has been stressed that central-exam systems focus on students as the pivotal stakeholders in education (Hanushek 2002). The recent discussion on accountability systems in the United States tends to argue that schools should be the primary unit of accountability (Ladd 2001). In this paper, I will argue—and present supporting evidence—that this contrast between central-exam and school-

[2]The only explicit consequence attached to central exams is often the dependence of grade promotion or graduation on performance on the exam.

based accountability systems in terms of whose behavior should be targeted may be more apparent than real. By focusing incentives on student performance, central-exam systems alter the way all stakeholders in education behave. The evidence suggests that the changes induced in the behavior of teachers and schools may actually be more important than the changes induced in the behavior of students.

1.2 Outline of the Paper and Summary of Findings

Outline of the Paper. Section 2 lays the theoretical foundation of the analysis by discussing the specific features of central-exam systems as accountability devices that can facilitate the monitoring of performance in the school system. It details the various channels by which central exams may impact educational outcomes, focusing on how they change the incentives faced by students, teachers, schools, administrators, and parents. Section 3 presents the international micro database derived from TIMSS and TIMSS-Repeat. It shows some descriptive statistics on the school systems in the participating countries and discusses the general opportunities and limitations of using international evidence to broaden our knowledge of how school systems function. Section 4 compares student performance in systems with and without central exams, including results by grade level, by performance quartile, and by family background as well as robustness checks for potential omitted-variable bias. In Section 5, the evidence goes into greater detail by showing how institutional features such as school autonomy, teacher influence, and parental involvement have different impacts in systems with and without central exams. This evidence sheds some light on how the behavior of the different educational stakeholders might be affected through central exams. Section 6 concludes by asking whether central exams lead to real increases in students' knowledge or merely to teaching and learning the test, and by some thoughts on the relative merits of central-exam systems in comparison to alternative accountability systems.

Summary of Findings. The evidence from TIMSS-Repeat confirms previous evidence from TIMSS that students in countries with central exit-exam systems perform better in their middle-school years both in math and in science than students in countries without central exams. This finding holds even after controlling for a large set of variables reflecting family background, resource endowment, and other institutional features of the school system. The size of the performance difference is substantial, lying in the range of 35 to 47 percent of an international standard deviation in test scores, and it increases from seventh to eighth grade. Students from each performance quartile of a country perform better in central-exam systems. While in math, higher-performing students seem to gain slightly more from central exams, no such difference is evident in science between performance quartiles. There is some evidence that central-exam systems dampen the effect of parental education on student performance, thereby compensating weak social backgrounds to some extent and leading to more equal educational opportunities for students from different social backgrounds. Evidence from including a large set of controls, particularly for the general centralization and other institutional features of the school systems and for the homogeneity of a country's population, and from restricting the analysis to within-region variation suggests that the case for substantial bias in the estimates is weak.

Central exams alter the way schools and teachers behave. Increased autonomy for schools in decision-making areas that include scope for opportunistic behavior, such as budgetary decisions and the determination of teacher salaries, has much more beneficial effects when central exams are in place. This finding is consistent with the claim that opportunistic behavior is decreased when central exams enable better monitoring of schools' behavior. At the same time, there is some evidence that central exams limit the useful freedom of schools and teachers in decision-making areas that do not include much scope for opportunistic behavior, such as day-to-day tasks like choice of supplies and textbooks. Central exams seem to ensure that student learning is focused on the educational goals of

the system, with the effect of regular testing and homework on student performance generally being more beneficial in central-exam systems. The involvement of interested parents is conducive to student performance in central-exam systems even when teachers deem this involvement as limiting their teaching, a result not found in systems without central exams. This may be attributable to the better information available to parents in central-exam systems.

Given that the performance tests of TIMSS and TIMSS-Repeat are general tests of students' knowledge in math and science on which representatives from all participating countries have agreed, these tests may be viewed as an independent test of whether central exams lead to real increases in the students' knowledge or whether they just lead to teaching and learning to the specific high-stakes central exam. The fact that students in central-exam systems perform better on the TIMSS and TIMSS-Repeat tests suggests that students in central-exam systems do indeed learn more in terms of mathematical and scientific knowledge, rather than just learning the specific central exam. Given the serious shortcomings of most school- and teacher-based accountability systems, central-exam systems seem to be a promising alternative device for focusing the incentives of all educational stakeholders on student learning.

2 How Central Exams Change Behavior

2.1 Central Exams as an Accountability Device to Mitigate Agency Problems

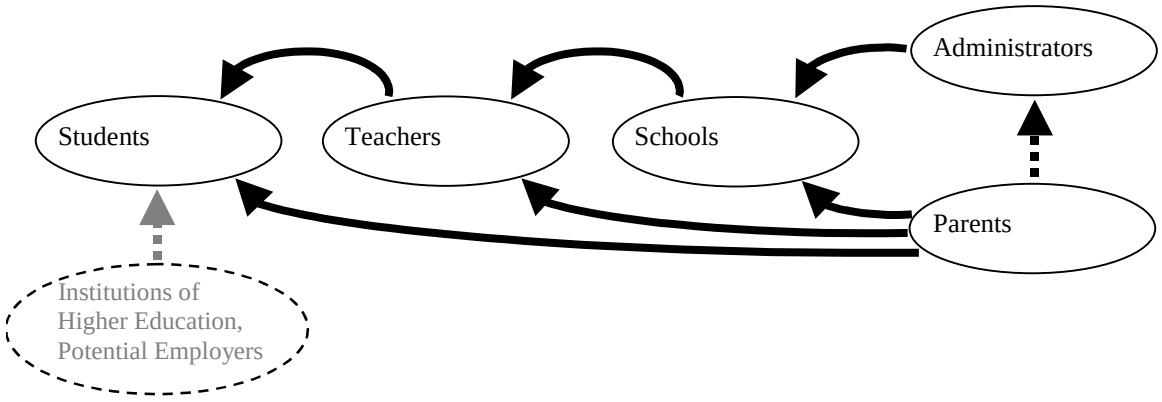
Accountability systems are often defined narrowly as systems that "reward and punish schools by allocating funding according to whether the school meets certain performance criteria" (Figlio and Page 2002). In this paper, I define accountability systems more broadly as any device that attaches consequences to measured educational performance. That is, the two common features of accountability systems are that they measure students' educational achievement directly, and that they attach consequences to measured performance. These consequences may be positive (rewards) or negative (sanctions), they may be implicit as well as financial or otherwise explicit, and their target may be any educational stakeholder, be it districts, schools, teachers, or students. While good performance would generally be rewarded, poor performance may lead either to sanctions or to more positive consequences such as additional assistance. Proponents of accountability systems hope that attaching consequences to student outcomes will lead to better educational performance in the school system (cf. Hanushek and Raymond 2001). Without such proper consequences, the motivation of educational stakeholders to put effort into improving educational outcomes may be rather low. By holding stakeholders accountable for performance, their incentives to work in order to yield superior performance are increased.

Why should explicit systems to introduce accountability be necessary in the first place? The answer is that a whole network of principal-agent relationships prevents accountability from being automatically secured in education. The first feature of a principal-agent relationship is "asymmetric information": The agent who is under a contract to a principal to perform a task has more information on what exactly he is doing than the principal. That is, the principal's monitoring of the agent's behavior is imperfect, limiting the principal's ability to hold the agent accountable. For example, teachers and parents do not perfectly know how much effort a student puts into learning; heads of school and parents cannot perfectly monitor how well a teacher prepares his or her lessons and what he or she does in the classroom. The second feature of a principal-agent relationship is that the agent and the principal have different interests. For example, students may be more interested in leisure relative to putting effort into learning than their parents would want them to be; heads of schools and teachers

may be more interested in their own finances and in a bearable workload relative to students' learning than parents and administrators would want them to be. Such differing interests make the lack of accountability in principal-agent relationships a problem. The extent to which agents' interests differ from their principals' interests will obviously depend on the specific task or decision-making area in question. For example, the difference in interests between principals and agents may be larger when the decisions affect the financial well-being of the agent than when they do not. Together, incomplete monitoring due to asymmetric information and divergent interests lead to the possibility of "opportunistic" behavior on part of the agent—that is, the agent will further his own interests rather than the principal's.

Accountability systems produce information on performance. As a result, they may be able to ease the monitoring problem inherent in principal-agent relationships. Central exams are one such accountability device. By producing comparable information on student performance, they go some way towards eliminating the informational asymmetry between principals and agents ubiquitous in education. Thus, they enable an improved monitoring of the behavior of the different stakeholders in education. Figure 1 provides a stylized picture of educational stakeholders and the monitoring relationships between them.[3] Students, the ultimate focus of the whole system, are most directly monitored by their parents and by their teachers. After having completed their general education, their educational performance may also be monitored by potential employers and institutions of higher education. Teachers are monitored by the heads of their schools and by the parents of the students whom they teach. Schools in turn are monitored by the educational administration and by the parents of their students.[4]

Figure 1: Monitoring in the School System



Central exams can provide the principals in this network of agency relationships with information that is not available in education systems without central exams, facilitating the monitoring of agents' behavior. The different principals may use this information in order to infer consequences on their agents in response to the agents' performance. This helps align the incentives of the different agents with the goal of the education system, namely the educational performance of the students. As a result, agents' effort to improve performance should increase, and teaching and learning should become more focused on the educational goals of the system. These beneficial effects of central exams on edu-

[3]This picture is rather stylized as further educational stakeholders and relationships can be thought of from which it abstracts. However, it is assumed that it identifies the most important features affecting student performance.

[4]In prolonging this chain of monitoring relationships, the educational administration might be viewed as being monitored by parents in its constituency and by the government, which in turn is monitored by the electorate, which comprises parents. In a sense, parents might be seen as coming nearest to something like an "ultimate" principal in this network of principal-agent relationships in education.

cational performance should be especially large when tasks are involved that include a large potential for opportunistic behavior, that is, when both informational asymmetries and interest differentials involved in the principal-agent relationship are large. In such cases, a lot of opportunism—diversion of behavior from the goals of the education system—can be curbed.

2.2 The Impact Channels of Central Exams: Effects on the Behavior of Parents, Administrators, Schools, Teachers, and Students

This basic mechanism of how central exams affect behavior in the education system can be detailed more clearly when focusing on the different principal-agent relationships depicted in Figure 1. The behavior of all educational stakeholders—parents, administrators, schools, teachers, and students—is affected by the existence of central exams, establishing several channels through which central exams may impact how school systems work and, ultimately, how students perform in terms of educational knowledge. In addition, the existence of central exams may change the way in which other institutional features of the school systems, such as the degree of decentralization in decision-making, affect behavior and student performance.[5]

Parents. Given central exams, parents have information on the performance of their children against an established standard and relative to other students in the education system. This is valuable information: Parents cannot only assess their child's performance against an absolute standard, but they also have some knowledge to decide on who might be responsible for this performance. For example, parents will generally know the performance of some other students in their child's class and the average performance in the country. Thus, in contrast to a system of teacher grading, parents now know whether it is mainly their own child who is doing badly or whether it is the whole class which is performing badly. That is, with central exams they are in a better position to monitor the performance of students, teachers, and schools. Consequently, parents are able to put pressure on students and/or teachers—whomever they deem responsible for the poor performance of their child.[6] When teachers grade their students themselves and students get marks relative to their class mean only, parents are not able to observe the performance of the class relative to the country mean and thus have no information on which to base a potential intervention. The existence of the information disseminated by the central exams on part of the parents is thus likely to affect the behavior of both students and teachers (see below). In the same way, parents can now monitor the performance of the whole school relative to other schools, and of the administrative entity relative to others. Moreover, the rather implicit monitoring by parents may have the advantage over any system of explicit monitoring by some administrative mechanism that, given their decentralized knowledge, parents may be able to assess fairly well what quality the student intake of a school has in terms of prior ability and thus what might and might not be expected in terms of ultimate performance. Explicit systems of school-based monitoring from above, in contrast, seem to be hard to implement in a meaningful way (see Section 6 below).

While central exams provide information to all parents, not necessarily all parents will be willing and able to make use of it. Thus, the impact of central exams might differ depending on how strongly parents care for their child's progress. In central-exam systems, parents who show interest in how

---

[5]For an incorporation of some of these impact channels of central exams into a simple formal model of educational production, see Bishop and Wößmann (2001). For further theoretical hypotheses on the effects of central-exam systems, see also Bishop (1995, 1999a).

[6]When central exams are only administered as exit exams at the end of secondary education, parents cannot directly monitor the performance of their child during the whole school career; however, central exit exams do tend to generate information inside the system on the relative performance of teachers, thereby still allowing parents to monitor teachers and thus improve teachers' incentives to further student performance.

much their child is learning have a meaningful foundation to intervene and will probably use this opportunity to pressure students and teachers to increase their effort, but parents who are less concerned with their child's educational performance may not make use of the additional information. Involvement of interested parents in the teaching process may thus be more beneficial in a central-exam system, while this channel may not work with parents who are less interested.

Administrators. Just like parents, administrators can also get valuable information from central-exam systems that enables them to monitor schools' performance and to draw consequences from their relative performance. While a school's low teaching effort might not be noted by administrators who lack comparable performance information, it would be more likely to attract administrators' attention with central exams. Administrators will usually have even more comparative information on performance than parents because they have access to measured performance for all the students in the systems and for successive years. This enables them to monitor the relative performance of schools and teachers even closer. Crucially, administrators will also have stronger incentives themselves to ensure good performance of the school system because, given the information spread by central exams, their behavior will now be more closely monitored by parents, the electorate, and the government.

Schools (Heads of School)[7] *and Teachers*. As central exams allow parents and administrators to monitor the performance of schools more closely, schools' behavior—usually expressed by the behavior of the heads of schools—should adapt correspondingly. Their incentives get centered on the educational performance of their students, and their leeway to act opportunistically—for example by using resources for usages that do not substantially further student achievement—is reduced. It can thus be expected that schools increase their effort to further student achievement and that they focus their work more closely on student achievement relative to other tasks.

Teachers' behavior should adapt equivalently, because just like schools, teachers are agents in a contract to teach the students. The main principals in these principal-agent relationships shift from administrators and parents in the case of schools to heads of schools and parents in the case of teachers. By producing information that facilitates the monitoring of agents' behavior, central-exam systems make sure that teachers must expect to face consequences for what they are doing. Thus, teachers get monitored and pressured to perform better by parents and by the heads of their schools, with an additional indirect impact at work in the latter case as the heads of schools themselves get pressured by parents. As a result, teachers' incentives are aligned more closely with student performance, leading to increased teacher effort and to a closer focus on the subjects covered by the central exams. Insofar as the central exams are designed to cover the standards that the education system is meant to pursue, this refocusing of efforts should be beneficial to students' knowledge in these areas. Thus, regular testing and homework assignment may get better focused on the core knowledge meant to be taught.

Central Exams and Autonomy of Schools and Teachers. The changes in incentives also have implications for how other features of the school system affect educational outcomes. Most importantly, the decentralization of decision-making should have different impacts in systems with and without central exams. Specifically, when schools are autonomous, they have ample leeway in their behavior. Whenever there is large room for opportunism on a decision—that is, when information asymmetries as well as differences in interests between schools and parents or administrators are both large—the extent of monitoring is vital to whether autonomous decisions will be carried out in the interest of student learning or not. Without central exams, schools with substantial autonomy may act in ways inconsistent with furthering student achievement without penalty, as their detrimental behavior cannot

---

[7]Given the equivocal meaning of the term "principal," this paper refers to the person responsible for a school as the "head of school." In terms of the principal-agent theory detailed above, the head of school is both an agent in his relationship with administrators and parents and a principal in his relationship with teachers.

be observed. With central exams, by contrast, the results of such opportunistic behavior will be observed, forcing schools to lean more towards behavior conducive to student performance.

Informational asymmetries are quite large in most areas of educational decision-making. However, the extent to which schools' own interests run counter to the interest of furthering student knowledge will depend on the specific task, or area of decision-making, in question. It might be expected that schools have a strong self-interest running counter to student learning whenever there is money involved in the decision, as it is only natural to try to increase the personal payoff for a given level of work (or, conversely, to reduce the level of work for a given payoff). It is in this group of tasks where devices that hold agents accountable should have their largest beneficial impact. By contrast, schools' own interests may be well in line with student learning in such decision-making areas as the choice of textbooks or supplies, as it is not obviously in the interest of schools to use poor supplies. In these tasks, the scope for opportunistic behavior is limited, and the need for accountability systems is correspondingly small.

The effects of school autonomy also depend on whether decentralized knowledge is important for a specific task. In many decision-making areas, local decision-making is likely to be more informed because it can draw on the decentralized knowledge that is available to schools, but not to any central entity. The extent to which decentralized knowledge is important again depends on the decision-making area in question. For example, the best way to transfer educational contents to students may vary among schools in different locations, making local discretion regarding the most suitable teaching techniques and supporting equipment vital. By contrast, decisions concerning the body of knowledge that students should be taught may be best made at a central level, rather than by individual schools.

These two considerations, severity of opportunism and importance of local knowledge, jointly determine the expected net impact of school autonomy in a given area of decision-making on students' educational achievement in systems without and with central exams. Table 1 summarizes these relationships. If, on the one hand, there is no danger of opportunism in a task, the impact of school autonomy will be equivalent in systems without and with central exams: It will have no impact on student performance when there is no specific local knowledge involved (cell [N1] in Table 1),[8] and it will be conducive to student learning when local knowledge is important to the task [N2a]. An example of the latter case may be the choice of specific teaching techniques, where local knowledge may be substantial and opportunistic interests limited. The one exception to the equivalence between the two systems in the absence of opportunism is when central exams limit the leeway within which schools can decide and when at the same time local behavior beyond this leeway would be superior [N2b]; in this case, central exams might reduce the extent to which school autonomy is beneficial for achievement.

If, on the other hand, the potential for opportunistic behavior is large, decentralized decision-making will have substantially different effects in systems without and with central exams. If there are no central exams and local knowledge is not vital [O1], the possibility of opportunistic behavior will make school autonomy strongly detrimental to student learning. As schools' incentives are focused on student learning when central exams are in place, the negative impact of local behavior will be eliminated. For example, schools' knowledge in budgetary matters may not be superior to the knowledge of external agencies, and school autonomy over their own budget may lead to opportunistic behavior and thus inferior student performance. The informational function of central exams may hinder detrimental opportunistic budgetary decisions of schools, eliminating the negative effects of budgetary autonomy.

---

[8] If central knowledge is superior in a task, then rendering schools autonomy in this task might even be detrimental to student achievement.

Table 1: The Impact of School Autonomy on Achievement Without and With Central Exams

Local knowledge \ Opportunism		Severe		No	
		without central exams	with central exams	without central exams	with central exams
Not important		[O1] – –	0	[N1] 0	0
Important	within limits set by central exams	[O2a] –	+ +	+	+
	beyond limits set by central exams	[O2b] –	+	+ +	+

– – = very detrimental; – = detrimental; 0 = no impact; + = conducive; + + = very conducive

When local knowledge is important for a task [O2], the negative impact of school autonomy in systems without central exams is lessened to some extent, although the detrimental effect of opportunism may still overcompensate the positive influence of local knowledge. But once opportunism is curbed through central exams, it can be expected that decentralized decision-making will lead to superior outcomes as it can draw on superior local knowledge. An example of such a situation may be the determination of teacher salaries: While there is scope for opportunism on part of the schools in this decision-making area, schools may have an informational advantage over external agencies about how well different teachers are performing. Without central exams, schools may not have much incentive to use this local knowledge in order to further student performance, so that opportunistic behavior may lead to negative effects of school autonomy in salary decisions; with central exams, schools have an inventive to use their local knowledge to improve student performance, and salary autonomy may thus be beneficial for student learning in central-exam systems. In short, changing the way in which decentralization affects outcomes is one impact channel through which central exams may affect educational performance, and they should be especially helpful whenever opportunism can be curbed.

Similarly to school autonomy, decentralization of decision-making authority to individual teachers will have different effects in systems with and without central exams. Teacher autonomy in tasks where local teacher knowledge might help informed decisions but where the scope for opportunism is substantial should change from being detrimental to student learning without central exams to being conducive under a central-exam system. An example of substantial scope for opportunistic behavior by teachers may be tasks involving money for supplies.

The scope for opportunistic behavior is also substantial when teachers as a group influence what is taught in class. Interest groups are generally formed to advance a group's interests relative to the interests advanced by other groups. In education, this may mean that teacher interest groups will lean towards furthering teachers' own interests over the interest of furthering students' knowledge. When teachers have group influence at the school level, central exams may again work as a device that focuses their incentives on academic achievement: Instead of watering down the curriculum taught in the school, group influence of school teachers may actually focus on improving teaching methods in a system of central exams which monitors their behavior and externally sets the contents that are meant to be covered in class. However, if teachers form interest groups at the central level where the central exams are set, the existence of a central-exam system may actually strengthen the negative effects of the actions of teacher groups, such as country-wide teacher unions, as these now can easily influence the standards of the whole education system.[9] Thereby, central-exam systems may be more susceptible

[9]Compare Evers' (2001) account of how attempts to introduce effective accountability measures in the United States got watered down by interest-group pressures of teacher unions.

to teacher unions' furthering of teachers' idiosyncratic interests over the interest of student achievement.

Students. Central-exam systems also align the incentives of students with increased educational achievement through several channels. Teachers and parents have better performance information to monitor students' behavior. As teachers' incentives are aligned more closely with students' educational performance through central exams (see above), their capability to monitor their students' performance enables and impels them to initiate appropriate consequences. Even more, parents can monitor their children's performance better given central exams, and it is generally the assessment and behavior by their parents that children care most about. Having to expect decisive actions as a result of their performance should change students' own effort to achieve high performance. A further channel through which central exams prompt students to achieve higher is by increased external rewards for learning. As potential employers and institutions of higher education have central exams at their disposal to assess applicants' educational performance, they can base their hiring decisions more on observed educational performance. Thereby, students get incentives from outside the school system to increase their performance.

Additionally, central-exam systems might alter the behavior of students' peers relative to a system of teacher grading. When teachers grade relative to the class level, peer pressure against learning ("nerd harassment") might be a viable strategy to lower average performance of the class, which allows every student in the class to get the same grades at a lower effort level (Bishop 1999a, 1999b). The existence of central exams should decrease this peer pressure against learning, because the mark received by one student is no longer affected by the marks of other students in the class and because lowering the standards taught in the class will hurt all students in the class.

It is sometimes hypothesized whether central-exam systems or other accountability systems affect students of different ability levels differently. On the one hand, if the standards to be tested are set too high, central exams might affect the behavior of high-performing students, but not of low-performing ones. While the impact channels running through altered teacher and school behavior might still render positive effects also for low-performing students, high performers would gain disproportionately. On the other hand, if the central exams were only minimum-competency tests, low-performing students might be positively affected, but top-performing ones might not be affected at all. A similar pattern would emerge if high-performing students were entirely self-motivated to achieve higher performance so that additional incentives might have no noticeable effects, while poor-performing students need some pressure from the incentives established by central exams. Contrarily again, high performers might get positively motivated by central exams, while poor performers might have their initiative blocked by the fear of not doing well enough. In the end, if the central exams are implemented to grade students into multiple levels of achievement, students of different ability levels might just be affected equivalently, and there might be no notable difference in the impact of central exams on the performance of students with different initial ability, as everybody responds to incentives. In a similar way, it is not clear ex ante whether students from different family backgrounds would be affected differently by central exams.

In sum, student performance may be expected to increase under a system of central exams. However, these improvements need not predominantly come directly through increased student effort, but may instead arise from many different indirect channels. Through increased monitoring, the incentives of all agents in education are directed at furthering students' knowledge. However, it is ultimately an empirical question how strongly the different stakeholders respond to their altered incentives by focusing their behavior on the advancement of student performance. Thus, the remainder of the paper analyzes the empirical evidence on the overall impact of central exams and on the behavioral responses of the different actors.

3 International Data

3.1 The Micro Databases of TIMSS and TIMSS-Repeat

While the mode of exam systems does vary within a few countries, the main variation is across countries.[10] Therefore, the empirical evidence in this paper draws on two large international comparative studies of student achievement, the Third International Mathematics and Science Study (TIMSS) of 1995 and its replication in 1999. In the following, the original TIMSS study will be referred to as TIMSS-95, and the repeat study as TIMSS-Repeat. While students from three different age levels were tested in TIMSS-95, the number of participating countries was by far the largest at the lower-secondary or middle-school level. The target population of TIMSS-95 in middle school were the two adjacent grades with the highest share of 13-year-old students, which were seventh and eighth grade in most countries. TIMSS-Repeat was conducted only at the middle-school level, with the target population being the upper grade of the two adjacent grades with the highest share of 13-year-old students (eighth grade in most countries). Within each participating country, a random sample of schools was selected, and one class within each target grade of these schools was randomly chosen and entirely tested in both math and science, yielding a representative sample of students within each country.

Both studies were conducted by the International Association for the Evaluation of Educational Achievement (IEA), an independent cooperation of national research institutes and governmental research agencies. The development of the test contents was a cooperative process involving national research coordinators from all participating countries. This, together with the fact that all participating countries endorsed the curriculum framework and that substantial efforts were made to ensure high-quality sampling and testing in all countries, should make the student performance tested in the TIMSS tests comparable across countries. As two-thirds of the test items of TIMSS-95 had been released to the public after the study was conducted, these items had to be replaced in TIMSS-Repeat. The items substituted were similar in terms of content, format, and level of difficulty. In both studies, a quarter of the items (meant to cover a third of the testing time) were free-response items, sometimes requiring extensive responses, while the remainder of the items were multiple-choice questions. Both studies also performed a test-curriculum matching analysis that restricted the analysis to items definitely covered in each country's curriculum; this had little effect on the overall achievement patterns.[11]

In addition to testing students in math and science, the two studies collected contextual information in three background questionnaires: a student questionnaire, a teacher questionnaire, and a school questionnaire. Each student answered questions about his or her demographic characteristics and home background. The math and science teachers of each tested class answered questions about their personal characteristics and classroom environments. Heads of school answered more general questions about the school's administrative structure.

The set of participating countries differed between the two studies. Of the 39 countries for which complete datasets had been available for TIMSS-95, 16 did not repeat the assessment in 1999. Thus, 15 of the 38 countries participating in TIMSS-Repeat were new to the international assessment. The difference in participating countries allows for a test of the robustness of previous findings obtained using TIMSS-95 data on a substantially altered set of countries.

---

[10] For a bivariate analysis of the cross-state variation within Germany of student performance in the final year of upper secondary school, see Baumert et al. (1999).

[11] For details on the content areas covered in the math and science tests, on sampling and implementation procedures, questionnaire development, translation verification tests, data collection, quality control procedures in all steps of the study, data processing, and test-score scaling in TIMSS-Repeat, see the TIMSS documentation contained in Mullis et al. (2000), Martin et al. (2000a, 2000b), and Gonzalez and Miles (2001). For details on TIMSS-95, see Wößmann (2002a) and the references therein.

Table 2 shows the countries participating in TIMSS-95 and in TIMSS-Repeat. The first two columns report the size of the student samples in each country in the TIMSS-95 and the TIMSS-Repeat assessments, respectively. The average sample size across all participating countries was 6,834 students in TIMSS-95 and 4,751 students in TIMSS-Repeat. In total, the TIMSS-95 database contains micro-level information on 266,545 individual students in seventh and eighth grade from 6,107 schools. The TIMSS-Repeat database contains equivalent information on 180,544 individual students in eighth grade from 6,068 schools. The subsequent columns of Table 2 report the average performance in math and science of the countries participating in TIMSS-95 and TIMSS-Repeat. For most of the countries that participated in both studies, the difference between the performance levels achieved in 1995 and 1999 was small and statistically insignificant (see Mullis et al. 2000; Martin et al. 2000b).

The TIMSS-95 database used in this paper is taken from Wößmann (2002a), which contains a detailed description of its construction and content. This database combines the TIMSS-95 performance data in math and science with data from the different background questionnaires for each individual student and includes imputed values for missing values of questionnaire data. The TIMSS-Repeat database was constructed for the purposes of this paper. The construction of this new database is described in the Appendix. The two micro databases based on the two TIMSS studies include rich student-level data for representative samples of students from all the participating countries. Drawing from the background-questionnaire data contained in the databases, the analysis in this paper uses 17 variables to control for students' family background, 13 variables to control for resource endowment and teacher characteristics, and 18 variables to control for the institutional setting of the education system.[12] To enable an even higher statistical precision in the estimation, the two databases are also pooled into one large TIMSS dataset containing information on 447,089 students.

The TIMSS micro databases were merged with data on the existence of central-exam systems in the participating countries. While in some countries central exams exist at several grade levels during secondary school, the most common form of central-exam systems is school-leaving exams at the end of the upper-secondary school level. Therefore, the measure of central exams used in this paper is whether a country (or region within a country) has a system of central exit exams or not, with all forms of "curriculum-based external exit exam systems" (CBEEES, see Bishop 1999a) included. The measure does not recognize university entrance exams, as these are usually not taken by all students and do not constitute an integrated part of the school system. The exam data used in this paper, most of which was provided by John Bishop, is based on reviews of comparative-education studies and educational encyclopedia, interviews with representatives of the national education systems, government documents, and background papers.[13] The data is presented in the final columns of Table 2. When central-exam systems were present in some parts of a country but not in others, the value indicates the share of students in the country facing central exams.[14]

While the central-exam data refers to exit exams at the end of upper-secondary school, the data on students' educational performance refers to the lower-secondary or middle-school level. Central exams might be expected to have the most direct impact on performance in the year leading to the exam, but their impact should also extend into lower levels. This is especially the case for exit exams, which tend to test all the knowledge learned in secondary school and whose signaling effect may change incentives during the whole school life of a student. As the impact of school-leaving exams should become more salient the closer students are to taking them, the effects on student performance should become stronger in higher grades. This implication can be tested using the TIMSS-95 data, which allows a comparison between seventh- and eighth-grade performance.

---

[12] For a complete list of these control variables, see Appendix Table A1a.

[13] For more details on the definition and characteristics of CBEEES, see Bishop (1997, 1999a).

[14] The lack of regional identifiers in the TIMSS database precludes a sub-national matching of central exams to students in these countries.

Table 2: Descriptive Statistics: Number of Students, TIMSS Test Scores, and Central Exams

	Students		TIMSS-95				TIMSS-Repeat		Central Exams	
			7th Grade		8th Grade		8th Grade			
	TIMSS-95	TIMSS-R.	Math	Science	Math	Science	Math	Science	Math	Science
Australia	12812	4018	498	504	530	545	525	540	0.81	0.81
Austria	5698	–	509	519	539	557	–	–	0	0
Belgium (Fl.)	5662	5259	558	529	565	550	557	534	0	0
Belgium (Fr.)	4849	–	507	442	527	471	–	–	0	0
Bulgaria	–	3272	–	–	–	–	510	518	1	1
Canada	16572	8770	494	499	527	531	531	534	0.51	0.51
Chile	–	5907	–	–	–	–	394	421	0	0
Colombia	5299	–	369	388	385	411	–	–	0	0
Cyprus	5827	3116	446	420	474	463	477	462	0	0
Czech Rep.	6671	3453	523	533	564	574	520	537	1	1
Denmark	4354	–	465	439	502	478	–	–	1	1
England	3538	2916	477	513	506	553	496	542	1	1
Finland	–	2920	–	–	–	–	520	537	1	1
France	5898	–	492	452	538	498	–	–	1	0
Germany	5744	–	485	500	509	531	–	–	0.35	0.35
Greece	7921	–	440	449	484	497	–	–	0	0
Hong Kong	6745	5179	564	495	588	522	582	529	1	1
Hungary	5978	3183	502	518	537	554	531	554	1	1
Iceland	3727	–	459	462	487	494	–	–	1	0
Indonesia	–	5848	–	–	–	–	401	433	1	1
Iran	7416	5301	401	436	428	470	423	448	1	1
Ireland	6201	–	500	495	527	538	–	–	1	1
Israel	1403	4193	–	–	522	525	466	470	1	1
Italy	–	3328	–	–	–	–	479	495	1	1
Japan	10271	4745	571	531	605	571	579	551	1	1
Jordan	–	5052	–	–	–	–	428	451	1	1
Korea, Rep.	5827	6114	577	535	607	565	587	551	1	1
Kuwait	1645	–	–	–	392	430	–	–	0	0
Latvia	4960	2845	462	435	494	485	504	500	0.50	0.50
Lithuania	5053	2361	428	403	477	476	481	486	1	1
Macedonia	–	4023	–	–	–	–	447	458	0	0
Malaysia	–	5577	–	–	–	–	519	492	1	1
Moldova	–	3711	–	–	–	–	468	458	1	1
Morocco	–	5402	–	–	–	–	336	319	1	1
Netherlands	4076	2943	516	517	541	560	538	544	1	1
New Zealand	6866	3613	472	481	508	525	491	511	1	1
Norway	5732	–	461	483	503	527	–	–	1	0.30
Philippines	–	6601	–	–	–	–	349	344	0	0
Portugal	6753	–	423	428	454	480	–	–	0	0
Romania	7471	3425	454	452	482	486	471	472	1	0
Russian Fed.	8160	4332	501	484	535	538	526	530	1	1
Scotland	5666	–	463	468	499	518	–	–	1	1
Singapore	8285	4966	601	545	643	607	603	568	1	1
Slovak Rep.	7101	3492	508	510	547	544	533	535	1	1
Slovenia	5603	3109	498	530	541	560	531	533	1	1
South Africa	–	8146	–	–	–	–	278	246	1	1
Spain	7595	–	448	477	487	517	–	–	0	0
Sweden	8855	–	477	488	519	535	–	–	0.50	0.50
Switzerland	11717	–	–	–	–	–	–	–	0	0
Taiwan	–	5772	–	–	–	–	584	572	1	1
Thailand	11627	5732	495	493	522	525	469	482	1	1
Tunisia	–	5051	–	–	–	–	446	428	1	1
Turkey	–	7841	–	–	–	–	430	432	1	1
United States	10967	9028	476	508	500	534	503	515	0.07	0.07

3.2 What Can Be Learned from International Evidence on Central Exams: Opportunities and Limitations

The use of international comparisons to estimate the effect of central exams on student performance presents both opportunities and limitations. Its main virtue lies in the fact that the institutional variation that exists between countries is an important source of information that can be exploited—institutional variation that is not given within most countries. Thus, the variation both in central exams and in the extent of school autonomy can help to shed light on the effects hypothesized in Section 2. To test the hypotheses on how central-exam systems affect student performance, variations in student performance have to be related to variations in exam systems and other institutional features. As these are not given within most countries, the international evidence has the potential to reveal relationships not usually evident in national data (see Wößmann 2002b; similarly, Hanushek 2002).

Understanding the sources of international variation in student performance is also an interesting research question in its own right. For example, recent research has shown that international differences in student performance matter a lot for international differences in economic growth and levels of development (Hanushek and Kimko 2000; Barro 2001; Wößmann 2003).

However, cross-country comparisons also face important limitations. First, the extent to which findings from cross-country evidence apply to individual countries may be limited. For example, if the research question is how a specific reform would affect performance in a specific country, it might be especially instructive to look at the performance in countries that are similar to the one in question in all respects except for the one regarding the reform issue. If, by contrast, the comparison country exhibits many other institutional features that differ from the country in question, assuming the same behavior and results in response to the reform might be poor inference, as different institutional settings may set different incentive environments and thus cause different behavioral responses (cf. Hoxby 2002b). This limitation can, however, be alleviated by using a multiple regression analysis that both incorporates multiple countries and controls for multiple influences. Controlling not only for the influences of family background and resource endowments, but also for institutional features of the school system, the most important effects of other features that might differ between countries should no longer affect the estimate of the specific reform issue of interest.

Even more, this is where the use of individual student data comes in as especially helpful. While much of the previous cross-country research was performed at the country level and was thus unable to account for differences in local features within the school systems (e.g., Bishop 1997; Lee and Barro 2001), the micro data used in this paper make it possible to look at the interactions of any reform issue, such as central-exam systems in this paper, with other local features (see Section 5 below). Thereby, the individual data can inform about how central exams work in different local settings, and a reader interested in probable effects within the surroundings of a specific school system can look at the effects within the particular settings of interest.

The most severe limitation facing a cross-country analysis of the impact of central exams probably is the potential for bias due to omitted variables. Among the features that vary across countries and that might in principle bias the coefficient on central exams in cross-country regressions, four country characteristics especially spring to mind: the overall degree of centralization of the education system; other institutional settings of the school systems; the homogeneity of a country's population; and cultural differences. Rather than being randomly distributed across countries, central exams may be more prevalent in generally centralized and homogenous countries, or in countries with other institutional and cultural characteristics in common. If these other characteristics affect educational outcomes, the estimates of the coefficient on central exams might reflect these other cross-country differences rather than the impact of central exams.

The extent to which central exams reflect the general centralization of a country or its school system can be checked by including control variables such as the percentage of public spending controlled by

the central government or the centralization of curriculum and textbook approval. The inclusion of controls for many additional institutional features of the school system allows an assessment of the potential for omitted-variable bias in the estimation of the impact of central exams due to these other institutional features. In the same way, it is possible to control for proxies for the homogeneity of a country's population. These specification tests should control for the most important biasing influences by other country features and incentive environments, thereby dampening the possibility of omitted-variables bias. A comparison of the estimates with and without the controls may also provide some indication of the potential size and direction of any bias (see Section 4.1 below).

In addition to potential biases due to centralization, institutions of the school systems, and homogeneity of the population, it is sometimes argued that much of the international variation in student performance may be due to more fundamental cultural differences. Insofar as such cultural differences are related to the existence of central-exam systems, the estimates of the coefficient on central exams will be biased. One possibility to assess the potential for omitted-variable bias from this direction is to include regional (continental) dummies as additional control variables. By controlling for any differences that might exist between regions, such an estimation considers only the within-region variation in central-exam systems and performance. As concerns about cultural differences generally arise in cross-continental comparisons—for example, in terms of Asian versus European values—but should not be substantial within world regions, estimates of the impact of central exams that control for regional differences should not be substantially biased by cultural differences. Furthermore, a comparison of the estimates with and without regional controls again allows for an evaluation of the potential size and direction of any bias (see Section 4.4 below).

Two more issues of the interpretation of the international evidence presented in this paper have to be addressed. First, it might be argued that the existence of central-exam systems may be endogenous to the level of educational performance in a country. This would again introduce bias into the estimates of the effect of central exams on student performance. However, it seems unlikely that endogeneity would introduce a noteworthy effect in this case, both because the potential size of such a bias may be deemed relatively small and because the existence of central-exam systems is generally a long-run institutional feature of the school systems that does not change often. As central-exam systems have been in place for decades in most countries, they would certainly not be endogenous to the performance of individual students in school today. Even more, the idea of endogeneity of a central-exam system would presumably be that governments introduce such a system in order to improve the poor performance of students. In this case, performance would have a negative effect on the prevalence of central exams, biasing standard least-squares estimates of the effect of central exams on student performance downwards. The estimates presented in this paper would thus be conservative estimates of the impact of central exams as they "err on the right side."

Second, there is unfortunately not much information about the specifics of the different central-exam systems in the different countries. In some systems, the performance of students and schools on the exam may be made publicly available, while this may not be the case in others. Some school systems may use the exam information to decide on whether a student is promoted to the next grade or has to repeat the grade, while others may not. Some systems may have regular central exams during secondary education, others not. Some central exams may be purely multiple choice, while others may have essay-type questions. Unfortunately, the evidence presented in this paper cannot say much about these specific features of different central-exam systems, but can only produce estimates of the general effect of whether there is central examination at all or not. As a further consequence of this and of the potential for the omission of other influence variables, the coefficients on central exams estimated in this paper should be interpreted as measures of the impact of central-exam systems and everything else that goes with them—which might, in some cases, be testing earlier in school, no-social-promotion policies, and other educational policies.

To sum up what can be learned from international comparisons, they establish a major source of information on the effects of central-exam systems not available in within-country research, as long as their limitations are borne in mind. If the research question is what role central exams can play in an explanation of the cross-country variation in student performance, attention has to be given to attempts to minimize the potential for biasing effects, and the interpretation of the results should bear these in mind. If the question is what a specific country—say, Germany or the United States—can learn from the international evidence on central exams and student performance, one should additionally focus the analysis on the effects of central exams in settings that are most relevant for the country. In Germany, the education system is relatively centralized at the regional level, with largely uniform funding and salary scales, and many decisions are rather bureaucratized. For example, the general rule is that schools do not have much say in the choice of their teachers, because teachers are selected by the regional bureaucracy and assigned to schools. Recent discussions on the relative merits of centralized versus decentralized systems may mean that the interaction effects of central exams with local autonomy may be especially informative for debates in Germany. In the United States, which has a highly decentralized school system with substantial local autonomy in terms of funding, teacher contracting, and curricular choices, the effects of central exams found in systems with high local autonomy seem especially relevant. While reduced-form estimates of the impact of central-exam systems may not necessarily translate directly to any specific country, a more detailed look at the different impact channels may nevertheless be highly informative for policymakers from all countries.

4 Central Exams and Student Performance: The International Evidence

4.1 Basic Results from TIMSS and TIMSS-Repeat

This section presents reduced-form estimates of the impact of central-exam systems on student performance, which reflect the total impact running through all conceivable impact channels—be it through altered behavior of parents, administrators, schools, teachers, or students. The coefficient of interest is the coefficient α on central exams in a regression of student performance on a host of explanatory variables:

$$(1) \qquad T_{ilsc} = \alpha E_c + B_{ilcs}\beta_1 + R_{lcs}\beta_2 + I_{lcs}\beta_3 + \eta_c + \nu_{sc} + \varepsilon_{ilsc}$$

where T_{ilsc} is the TIMSS math or science test score T of student i in class l in school s in country c. These test scores have been divided by the standard deviation of the test scores of all students in order to facilitate interpretation of coefficients and enable comparisons with other studies using different tests.[15] E_c denotes central exams, measured at the country level. B_{ilsc} is a vector of variables reflecting background characteristics of the student and his or her family, R_{ilsc} is a vector of measures of resource endowment and teacher characteristics, and I_{lsc} is a vector of variables depicting other institutional features of the school system such as the centralization of other features of the school system, school autonomy, teacher influence, or parental involvement. The latter two sets of variables are mostly measured at the classroom or school level. The error term has a country-level component η, a school-level component ν, and a student-level component ε.

This structure of the error term is implemented by using clustering-robust regression techniques that allow any degree of correlation among the error terms within each cluster in order to obtain consistent

[15] Note, however, that the standard deviation of test scores in an international setting may be larger than the standard deviation of scores on tests undertaken within individual countries.

estimates of standard errors in the presence of an hierarchical data structure (cf. Moulton 1986; Deaton 1997). For variables measured at the country level—like central exams—the standard errors reported in the tables use countries as the clustering unit, reflecting the fact that the number of independent observations on this variable is not the number of students, but only the number of countries. For all other variables, measured at the student, classroom, or school level, the standard errors reported in the tables use schools as the clustering unit, as schools constitute the primary sampling unit (PSU) in TIMSS. All regressions are weighted least-squares estimations that use the TIMSS sampling weight of each student as their weights. The weighting ensures that the proportional contribution of each stratum in the sample to the coefficient estimates is equal to the one that would have been obtained had there been a complete census enumeration (cf. DuMouchel and Duncan 1983), and it ensures that each country gets the same weight within the international estimation.

The results of the base regressions are reported in Table 3a for math, and in Table 3b for science. In both subjects, the results are presented separately for the TIMSS-95 test, for the TIMSS-Repeat test, and for the combined dataset that pools the results of both tests. The first estimates reported in each of the three blocs stem from an estimation that regresses student performance on central exams only, without any additional controls—that is, restricting the coefficients β_1, β_2, and β_3 of equation (1) to zero. The subsequent columns in each bloc successively add controls for family background, for resource endowment and teacher characteristics, and for institutional features of the school system.

Before comparing the estimates obtained with the different sets of controls, the focus is on the results presented in the last column of each bloc. With the most encompassing set of control variables, these estimates should come closest to the actual impact of central exams on student performance. The complete results of this last specification for the three samples are reported in Appendix Tables A1a and A1b for the two subjects.[16]

According to these estimates, students in countries with central-exam systems scored 40.9 percent of a standard deviation higher on the TIMSS-95 math test than students in countries without central-exam systems, controlling for effects of family, resource, and institutional background. Similarly, the lead of students in countries with central exams was 47.0 percent of a standard deviation in the TIMSS-Repeat math test. In the pooled math regression, the lead was 42.7 percent of a standard deviation. In science, students in countries with central-exam systems scored 39.7 percent of a standard deviation higher in TIMSS-95, 35.9 percent higher in TIMSS-Repeat, and 35.9 percent higher in the pooled analysis. All these coefficients on central exams are statistically significant at the 1 percent level.

Thus, the first result of this analysis is that the findings based on the TIMSS-Repeat test, with its differing set of participating countries, confirm previous findings derived from TIMSS-95 (Bishop 1997, 1999a; Wößmann 2002a) that central exams seem to exert a substantial positive impact on the educational performance of students. Furthermore, the size of the estimated coefficient is robust to the use of the new dataset: When the impact size is allowed to differ between the two tests in the pooled regression by including an interaction term between central exams and a dummy for the TIMSS-Repeat test, the difference in the size of the estimate is statistically insignificant both in math and in science.

The substantial size of this impact estimate can be seen when comparing it to the impact sizes estimated for other policy reforms in other studies. For example, Krueger (1999) found for the Tennessee Project STAR that reducing class size in primary schools by seven to eight students (from about 23 to about 16 students) led to an increase in test scores of about 0.22 standard deviations. This estimate of the impact of reduced class size on student performance is at the upper bound of what other studies have found, and some have argued that the design of the experiment might have biased

---

[16] In the pooled regressions, a control for the study year—TIMSS-95 versus TIMSS-Repeat—was never statistically significant and was consequently dropped from the estimations.

Table 3a: The Impact of Central Exams on Student Performance: Base Estimates (Math)[a]

	TIMSS-95				TIMSS-Repeat				Pooled			
Central exams	0.387 °	0.416 *	0.357 +	0.409 *	0.288	0.444 +	0.555 *	0.470 *	0.297 °	0.397 *	0.413 *	0.427 *
	(0.195)	*(0.150)*	*(0.147)*	*(0.135)*	*(0.281)*	*(0.171)*	*(0.146)*	*(0.135)*	*(0.163)*	*(0.114)*	*(0.111)*	*(0.098)*
Family controls [17]		✓	✓	✓		✓	✓	✓		✓	✓	✓
Resource/teacher controls [13]			✓	✓			✓	✓			✓	✓
Institutional controls [18]				✓				✓				✓
Students (unit of observation)	266545	266545	266545	266545	180544	180544	180544	180544	447089	447089	447089	447089
Countries	39	39	39	39	38	38	38	38	77	77	77	77
R^2	0.029	0.182	0.208	0.238	0.011	0.289	0.332	0.362	0.015	0.222	0.257	0.285

[a]Dependent variable: TIMSS international math test score. Clustering-robust standard errors in parentheses. All standard errors reported in this table take countries as the level of clustering. Significance levels (based on clustering-robust standard errors): * 1 percent, + 5 percent, ° 10 percent.

Table 3b: The Impact of Central Exams on Student Performance: Base Estimates (Science)[a]

	TIMSS-95				TIMSS-Repeat				Pooled			
Central exams	0.396 *	0.428 *	0.357 *	0.397 *	0.267	0.335 +	0.420 *	0.359 *	0.274 +	0.343 *	0.371 *	0.359 *
	(0.134)	*(0.113)*	*(0.109)*	*(0.099)*	*(0.232)*	*(0.152)*	*(0.121)*	*(0.129)*	*(0.129)*	*(0.097)*	*(0.088)*	*(0.083)*
Family controls [17]		✓	✓	✓		✓	✓	✓		✓	✓	✓
Resource/teacher controls [13]			✓	✓			✓	✓			✓	✓
Institutional controls [18]				✓				✓				✓
Students (unit of observation)	266545	266545	266545	266545	180544	180544	180544	180544	447089	447089	447089	447089
Countries	39	39	39	39	38	38	38	38	77	77	77	77
R^2	0.033	0.175	0.193	0.205	0.011	0.278	0.310	0.326	0.015	0.208	0.242	0.256

[a]Dependent variable: TIMSS international science test score. Clustering-robust standard errors in parentheses. All standard errors reported in this table take countries as the level of clustering. Significance levels (based on clustering-robust standard errors): * 1 percent, + 5 percent.

the estimate upwards (cf. Hanushek 1999; Hoxby 2000). But even when taking this estimate at face value, the estimated impact of central exams is two-thirds larger in science and twice as large in math. Furthermore, reducing class size by one-third would increase educational spending by one-third. By contrast, implementing a central-exam system seems to have a negligible effect on overall educational spending, a fact suggested by cross-country evidence[17] and supported by findings for the United States (Hoxby 2002a). Considering the size of their performance impact and their cost effectiveness, central-exam systems would seem an attractive policy alternative.

When comparing the estimates in Tables 3a and 3b obtained with different sets of controls to one another, the variation in point estimates is actually rather small compared to the size of the estimates. For example, the estimates for the TIMSS-95 test all lie within about half a standard error of one another—that is, none of the differences is statistically significant. In the pooled regressions, even the largest difference between two adjacent estimates, the math estimates without any controls versus with family controls, is not statistically significant. The null hypothesis of a Hausman test that the difference between these two coefficients is not systematic cannot be rejected.[18] While not being statistically significantly different from each other, the point estimates of the coefficient on central exams do increase slightly when controls for students' personal and family characteristics are included, implying that countries with more favorable family background are less likely to have central-exam systems. The slight increase in the TIMSS-Repeat point estimates due to the additional inclusion of resource and teacher controls dominates the slight decrease in the TIMSS-95 point estimates in the pooled estimation. While the opposite is true for the additional inclusion of institutional controls in science, the point estimate in the pooled math regression again slightly increases. In sum, while the omission of the measurable family-background controls seems to bias the estimate of the impact of central exams downwards, the direction of the bias due to the omission of measurable resource and institutional controls is less clear and its size is small.

It should be particularly noted that the set of institutional controls includes two measures of the general centralization of the school system, namely the centralization of the curriculum and of textbook approval. The coefficient estimates on these two controls are consistently positive, but much smaller than the estimate on central exams and often statistically insignificant (cf. Tables A1a and A1b). Their inclusion does not change the estimate on central exams substantially, suggesting that the latter does not primarily pick up effects of the general centralization of the school system.[19] Likewise, including a measure of ethnolinguistic fractionalization as a proxy for the homogeneity of a country's population leaves the substantive results on central exams unchanged.[20] These robustness tests suggest that any potential bias from omitted variables is rather small and generally tends to work in the downward direction.

It is informative to analyze how much of the international variation in student performance is due to the existence versus lack of central-exam systems in the different countries. In the specification of the pooled datasets that includes all the family, resource/teacher, and institutional control variables, adding the central-exam variable increases the explained proportion of the total variation in student

---

[17]The existence of a central-exam system does not have a statistically significant relationship with the cross-country distribution of per-student educational expenditure. The point estimate is slightly negative, suggesting that a gain in effectiveness of resource usage might even overcompensate any direct cost of implementing a central-exam system.

[18]The $\chi^2_{(1)}$ of the Hausman test is 0.73 (probability > χ^2 = 0.391).

[19]Using the share of educational funds controlled by the central level of government as an alternative control variable for the overall centralization of a country's school system (available for OECD countries from OECD 2000) also does not change any of the estimates on central exams in a noteworthy way. Actually, centralized financing has a statistically significant *negative* coefficient in most specifications.

[20]The measure of ethnolinguistic fractionalization is defined as the probability that two randomly selected persons from a given country will not belong to the same ethnolinguistic group (taken from the World Handbook of Political and Social Indicators as reported in Mauro 1995). It is never statistically significantly related to student performance in the regressions.

test scores (the R^2) by 2.4 percentage points in math (from 0.260 to 0.285) and by 1.9 percentage points in science (from 0.237 to 0.256). Relative to the total cross-country variation, which is 34.1 percent of the total cross-student variation in math test scores and 28.5 percent of the total variation in science test scores, this proportion of the variance additionally explained by central-exam systems is 7 percent of the total cross-country variation, both in math and in science. That is, about 7 percent of the international variation in math and science performance can be attributed to the existence of central-exam systems. In medical research, the variance in health status accounted for by treatment effects is often less than 1 or 2 percent. For example, the first studies associating smoking and longevity accounted for only around 2 percent of the variance in longevity (Berliner 1990). Yet the responses in policy and medical practice to such findings are tremendous.

4.2 Effects by Grade

The measure of central exams used in this paper is one of exit exams at the end of upper-secondary school, while student performance is tested in seventh and eighth grade. Thus, the reported results suggest that central exit-exam systems send incentive signals down to grades in lower-secondary school. As suggested in Section 3.1 above, these incentive signals might be expected to be stronger in eighth grade than in seventh grade. This hypothesis can be tested using the TIMSS-95 data, as students from both grade levels were tested in this test.[21] Unfortunately, this is not possible for the TIMSS-Repeat data, which tested only eighth-grade students.

Table 4 presents results on the interaction effect between central exams and grade level. The impact of central exams on TIMSS-95 math performance was 14.4 percent of a standard deviation larger in eighth grade than it was in seventh grade. Likewise, the impact of central exams on eighth-grade science performance was 8.0 percent of a standard deviation larger than their impact in seventh grade. Thus, the impact that central exit exams exert on student performance indeed seems to grow over the course of secondary education.

Table 4: The Impact of Central Exams on Student Performance by Grade[a]

	Math				Science			
	TIMSS-95		Pooled		TIMSS-95		Pooled	
	Coef.	*S.E.*	Coef.	*S.E.*	Coef.	*S.E.*	Coef.	*S.E.*
Central exams[b]	0.332 *	*(0.117)*	0.313 *	*(0.099)*	0.354 *	*(0.097)*	0.316 *	*(0.090)*
Central exams × upper grade	0.144 *	*(0.016)*	0.156 *	*(0.016)*	0.080 *	*(0.014)*	0.059 *	*(0.014)*
Upper grade	0.279 *	*(0.014)*	0.263 *	*(0.015)*	0.407 *	*(0.011)*	0.401 *	*(0.013)*
Family controls [16]	✓		✓		✓		✓	
Resource/teacher controls [13]	✓		✓		✓		✓	
Institutional controls [18]	✓		✓		✓		✓	

[a]Each column reports results from one regression. Dependent variable: TIMSS international math/science test score. Clustering-robust standard errors in parentheses. Standard errors have schools as the level of clustering unless noted otherwise. Significance levels (based on clustering-robust standard errors): * 1 percent. – [b]Standard error has countries as the level of clustering.

[21]Two countries—Sweden and Switzerland—also tested some students in ninth grade in TIMSS-95, but this does not give enough variation in central-exam systems to meaningfully test whether their impact is even stronger in ninth grade.

4.3 Effects by Performance Quartiles and Students' Background

Section 2.2 discussed whether the impact of central exams might differ between students of different ability levels and family backgrounds. Table 5 reports results on the coefficient on central exams by performance quartile. In Panel A, each row reports the results from a regression on a different sample of students. The first row shows the coefficient on central exams for the sample of students that form the bottom quartile in each country in terms of their performance on the respective TIMSS test, while the second row has the students that form the lower-middle performance quartile, and so on. Note, first, that central exams have a statistically significant and large positive effect on student performance in each quartile sample in each test. That is, both relatively poor-performing students and relatively high-performing students gain from the existence of a central-exam system.

In TIMSS-95, the impact seems to be larger for students in higher quartiles, while in TIMSS-Repeat it seems slightly smaller for students in higher quartiles. To see whether these differences in point estimates are statistically significant, Panel B of Table 5 reports interaction terms of central exams with the successive performance quartiles. None of the differences in the size of the impact of central exams between performance quartiles is statistically significant in the TIMSS-Repeat math or science test. By contrast, the pattern of an increase in the impact by performance quartile in the TIMSS-95 math and science tests is statistically significant. While this shines through to the pooled regression in math, the differences are not statistically significant in the pooled regression in science. Thus, there is some evidence that the knowledge gain that high-performing students reap from the existence of central exams is larger than the knowledge gain that poor-performing students reap in math, but probably not in science. Still, any of these differences is small relative to the general gain produced by central exams across students from all performance quartiles. The bottom line is that both poor- and high-performers perform substantially better under a central-exam system than under a system without central exams.

Table 6 reports results on differences in the effect of central exams for students with different family backgrounds by including interaction terms of central exams with the family-background variables. The columns labeled "Coefficient" report the coefficient estimate on the family-background variable itself, while the columns labeled "Interaction" report the coefficient estimate on the interaction term between the family-background variable and central exams of the same regression. Family background itself exerts strong effects on students' educational performance. Students perform better in both math and science if they were born in the country in which they are currently living and going to school, if their parents were born in that country, if their parents have higher educational attainment, and if there is a larger number of books in their home—the latter generally serving as a proxy for the educational and social background of the home in which the students are raised. All these effects are large and statistically significant.[22]

There are differences in the impact of central exams for students with different family backgrounds. First, central exams dampen the effect of the country of birth of students and their parents. That is, immigrants seem to gain more from central-exam systems than nationally born students. Second, central exams also decrease the effect of parental education. Under a system of central exams, it seems to matter less from which parental background a student comes. While these differences are statistically significant for TIMSS-95 in both math and science, for TIMSS-Repeat they are only statistically significant in math. Third, there is not much evidence that central exams affect students from homes with different amounts of books differently in math, but in the TIMSS-95 science study, the positive effect of having more books at home is larger in central-exam systems. This finding counters the effect for parental education, which goes in the opposite direction.

---

[22]The one exception to this statement is the effect of parents' country of birth in the TIMSS-Repeat study, where all its impact is captured by the effect of students' country of birth, two variables that are highly collinear.

Table 5: The Impact of Central Exams on Student Performance by Performance Quartiles[a]

	Math						Science					
	TIMSS-95		TIMSS-Repeat		Pooled		TIMSS-95		TIMSS-Repeat		Pooled	
	Coef.	*S.E.*	Coef.	*S.E.*	Coef.	*S.E.*	Coef.	*S.E.*	Coef.	*S.E.*	Coef.	*S.E.*
Panel A												
Bottom-quartile regression	0.304 *	*(0.109)*	0.462 *	*(0.137)*	0.364 *	*(0.097)*	0.318 *	*(0.095)*	0.392 +	*(0.152)*	0.344 *	*(0.100)*
Lower-middle-quartile regression	0.427 *	*(0.134)*	0.464 *	*(0.138)*	0.453 *	*(0.101)*	0.403 *	*(0.102)*	0.369 *	*(0.132)*	0.382 *	*(0.089)*
Upper-middle-quartile regression	0.484 *	*(0.151)*	0.448 *	*(0.134)*	0.481 *	*(0.102)*	0.443 *	*(0.112)*	0.325 *	*(0.119)*	0.384 *	*(0.082)*
Top-quartile regression	0.539 *	*(0.169)*	0.435 *	*(0.127)*	0.503 *	*(0.107)*	0.509 *	*(0.127)*	0.289 +	*(0.109)*	0.403 *	*(0.081)*
Panel B												
Central exams	0.301 *	*(0.109)*	0.457 *	*(0.143)*	0.365 *	*(0.097)*	0.338 *	*(0.099)*	0.395 +	*(0.158)*	0.358 *	*(0.097)*
Central exams × lower-middle quartile	0.117 +	*(0.044)*	-0.003	*(0.033)*	0.086 +	*(0.033)*	0.065 °	*(0.035)*	-0.033	*(0.049)*	0.031	*(0.029)*
Central exams × upper-middle quartile	0.183 +	*(0.074)*	0.003	*(0.054)*	0.119 +	*(0.052)*	0.103 °	*(0.056)*	-0.069	*(0.084)*	0.030	*(0.048)*
Central exams × top quartile	0.263 +	*(0.109)*	0.016	*(0.081)*	0.154 +	*(0.073)*	0.161 °	*(0.086)*	-0.078	*(0.125)*	0.037	*(0.072)*
Lower-middle quartile	0.650 *	*(0.035)*	0.712 *	*(0.030)*	0.650 *	*(0.030)*	0.781 *	*(0.026)*	0.759 *	*(0.047)*	0.748 *	*(0.026)*
Upper-middle quartile	1.158 *	*(0.061)*	1.171 *	*(0.052)*	1.128 *	*(0.047)*	1.335 *	*(0.043)*	1.253 *	*(0.082)*	1.264 *	*(0.043)*
Top quartile	1.936 *	*(0.092)*	1.810 *	*(0.077)*	1.836 *	*(0.066)*	2.156 *	*(0.066)*	1.923 *	*(0.121)*	2.010 *	*(0.063)*

[a]*Panel A*: Each cell reports results from an individual regression. In each row, the sample of students included in the regression is only one quartile of the students in each country in terms of their performance. *Panel B*: Each column reports results from one regression. Dependent variable: TIMSS international math/science test score. All regressions control for all the family, resource, and institutional control variables reported in Table A1a. Clustering-robust standard errors in parentheses. All standard errors reported in this table take countries as the level of clustering. Significance levels (based on clustering-robust standard errors): * 1 percent, + 5 percent, ° 10 percent.

Table 6: The Impact of Central Exams on Student Performance by Family Background[a]

	Math						Science					
	TIMSS-95		TIMSS-Repeat		Pooled		TIMSS-95		TIMSS-Repeat		Pooled	
	Coef.	Inter.	Coef.	Inter.	Coef.	Inter.	Coef.	Inter.	Coef.	Inter.	Coef.	Inter.
Central exams[b]	0.697 *		0.906 *		0.658 *		0.437 *		0.739 +		0.439 °	
	(0.194)		*(0.232)*		*(0.216)*		*(0.151)*		*(0.325)*		*(0.232)*	
Upper grade	0.293 *	0.126 *			0.246 *	0.182 *	0.411 *	0.074 *			0.394 *	0.071 *
	(0.014)	*(0.016)*			*(0.015)*	*(0.016)*	*(0.011)*	*(0.014)*			*(0.012)*	*(0.014)*
Born in country	0.091 *	-0.018	0.392 *	-0.242 *	0.211 *	-0.052 °	0.144 *	-0.068 +	0.495 *	-0.310 *	0.288 *	-0.113 *
	(0.018)	*(0.028)*	*(0.036)*	*(0.045)*	*(0.021)*	*(0.029)*	*(0.020)*	*(0.029)*	*(0.041)*	*(0.050)*	*(0.022)*	*(0.031)*
Parents born in country	0.147 *	-0.192 *	-0.016	-0.014	0.120 *	-0.176 *	0.172 *	-0.051	-0.001	-0.020	0.112 *	-0.093 *
	(0.025)	*(0.034)*	*(0.038)*	*(0.047)*	*(0.022)*	*(0.029)*	*(0.025)*	*(0.032)*	*(0.052)*	*(0.059)*	*(0.025)*	*(0.031)*
Parents' education												
Finished primary	0.395 *	-0.257 *	0.149 *	-0.078 +	0.272 *	-0.200 *	0.263 *	-0.142 *	0.147 *	-0.031	0.150 *	-0.096 *
	(0.018)	*(0.023)*	*(0.034)*	*(0.039)*	*(0.018)*	*(0.022)*	*(0.016)*	*(0.021)*	*(0.036)*	*(0.041)*	*(0.015)*	*(0.019)*
Secondary	0.389 *	-0.393 *	0.417 *	-0.198 *	0.356 *	-0.233 *	0.223 *	-0.196 *	0.349 *	-0.058	0.192 *	-0.037 °
	(0.021)	*(0.027)*	*(0.038)*	*(0.043)*	*(0.019)*	*(0.024)*	*(0.017)*	*(0.024)*	*(0.037)*	*(0.043)*	*(0.016)*	*(0.021)*
Finished university	0.538 *	-0.254 *	0.593 *	-0.182 *	0.502 *	-0.178 *	0.396 *	-0.087 *	0.522 *	-0.060	0.360 *	-0.014
	(0.023)	*(0.029)*	*(0.042)*	*(0.048)*	*(0.021)*	*(0.026)*	*(0.019)*	*(0.026)*	*(0.041)*	*(0.047)*	*(0.018)*	*(0.024)*
Books at home												
11-25	0.122 *	0.002	0.181 *	-0.053 +	0.167 *	-0.020	0.087 *	0.076 *	0.186 *	-0.039	0.149 *	0.028
	(0.032)	*(0.036)*	*(0.019)*	*(0.022)*	*(0.020)*	*(0.023)*	*(0.020)*	*(0.025)*	*(0.020)*	*(0.024)*	*(0.016)*	*(0.020)*
26-100	0.314 *	0.084 +	0.409 *	-0.057 +	0.384 *	0.038	0.266 *	0.139 *	0.426 *	-0.037	0.344 *	0.110 *
	(0.033)	*(0.039)*	*(0.024)*	*(0.027)*	*(0.022)*	*(0.026)*	*(0.022)*	*(0.027)*	*(0.023)*	*(0.028)*	*(0.017)*	*(0.022)*
101-200	0.471 *	0.028	0.514 *	-0.022	0.521 *	0.020	0.395 *	0.187 *	0.572 *	-0.019	0.465 *	0.162 *
	(0.034)	*(0.040)*	*(0.026)*	*(0.030)*	*(0.023)*	*(0.027)*	*(0.023)*	*(0.029)*	*(0.027)*	*(0.032)*	*(0.019)*	*(0.024)*
More than 200	0.544 *	0.035	0.595 *	-0.025	0.587 *	0.026	0.486 *	0.196 *	0.616 *	0.022	0.530 *	0.186 *
	(0.035)	*(0.040)*	*(0.028)*	*(0.033)*	*(0.024)*	*(0.027)*	*(0.024)*	*(0.030)*	*(0.028)*	*(0.034)*	*(0.019)*	*(0.025)*

[a]Every two columns headed "Coef." and "Inter." together report the results of one regression. The column headed "Coef." reports the coefficient on the variable labeled in each row, while the column headed "Inter." reports the coefficient on the interaction term between central exams and the variable labeled in the row. Dependent variable: TIMSS international math/science test score. All regressions control for all the family, resource, and institutional control variables reported in Table A1a. Clustering-robust standard errors in parentheses. Standard errors have schools as the level of clustering unless noted otherwise. Significance levels (based on clustering-robust standard errors): * 1 percent, + 5 percent, ° 10 percent. – [b]Standard error has countries as the level of clustering.

In sum, the results show that the disadvantage of coming from a less beneficial family background seems to be reduced by central exams in math, while the pattern is less clear in science. This suggests that, at least in math, central-exam systems work towards equalizing opportunities for students from different family backgrounds. Together, the evidence on effects by performance quartiles and by family background suggests that high-ability students from poor family backgrounds seem to gain the most from central exams.

4.4 The Potential for Bias in the Estimates

It has been shown in Section 4.1 that the biases from omitting measurable variables of family, resource, and institutional background are relatively small and generally attenuate the estimate of the impact of central exams. Likewise, neither controls for the general centralization of the school system nor for the homogeneity of a country's population change the results. These findings may dampen concerns about any potential bias due to unmeasured omitted variables.

As argued in Section 3.2, eliminating the inter-regional variation and confining the analysis to intra-regional variation might be another way to evaluate the potential for biases, as any biasing impact of differential cultural backgrounds should be mitigated. Table 7 presents results of including eight regional dummies. The residual category are Western European countries, and dummies are included for North America, South America, Eastern Europe, Oceania (Australia and New Zealand), Asia, Middle East, Northern Africa, and Southern Africa (which is a country dummy for South Africa). The first thing to note is that in each estimation, only between two and four of the eight coefficient estimates on the dummies are statistically significant. In no case is there a statistically significant difference in the performance of countries from Western Europe, North America, and Oceania, after controlling for central exams and for the family, resource, and institutional variables considered in this study. The performance difference to Eastern European countries is statistically significant only on the TIMSS-95 math test. Asian countries tend to perform better than Western European ones in math, but not in science. South African students perform worse, while this is true only in some cases and to a lesser extent in South America, the Middle East, and North Africa.

Comparing the coefficient on central exams in the regressions including all these regional dummies to the coefficient estimates in Tables 3a and 3b reveals that only in the TIMSS-95 math study, the positive coefficient estimate on central exams becomes small and statistically insignificant after including the regional dummies. In this case, the positive estimate in Table 3a seems to mostly come from between-regional variation, and it is not clear whether this estimate captures an actual impact of central exams or effects of other cross-regional differences that go with them. In the TIMSS-Repeat math study and in the TIMSS-95 science study, the estimate on central exams does not change much by the inclusion of the regional dummies, and in the TIMSS-Repeat science study, it increases. In the pooled regressions, a statistically significant positive estimate prevails which is smaller in the case of math when the regional dummies are included and larger in the case of science. Neither in the pooled math estimation nor in the pooled science estimation is the difference in the coefficient estimate on central exams between the regression with and without regional dummies statistically significant.[23] Thus, the case for substantial omitted-variable bias seems to be weak also on the basis of the comparison of the base estimation to the within-regional estimation. The cautious conclusion to be drawn from these findings should be that the potential size of a bias seems small in most cases, that the direction of any bias can be either upwards or downwards, and that the case against interpreting the

[23]Based on Hausman tests, the standard error of the difference of –0.142 in math is *0.089*, and the standard error of the difference of 0.058 in science is *0.069*. The only case where the difference is statistically significant is, obviously, the TIMSS-95 math case (0.348 *(0.093)*); this is also the only case where the Hausman test rejects the null hypothesis that all coefficients (including the control variables) combined are not systematically different ($\chi^2_{(45)}$ = 61.89, probability > χ^2 = 0.048).

Table 7: The Impact of Central Exams on Student Performance Controlling for Regional Dummies[a]

	TIMSS-95		TIMSS-Repeat		Pooled	
	Coef.	S.E.	Coef.	S.E.	Coef.	S.E.
			Math			
Central exams	0.061	*(0.097)*	0.477 $^{+}$	*(0.203)*	0.286 $^{+}$	*(0.132)*
North America	-0.147	*(0.159)*	-0.039	*(0.185)*	-0.147	*(0.111)*
South America	-0.598 *	*(0.128)*	-0.043	*(0.347)*	-0.074	*(0.163)*
Eastern Europe	0.383 *	*(0.129)*	0.303	*(0.216)*	0.479 *	*(0.113)*
Oceania	-0.044	*(0.123)*	-0.042	*(0.108)*	-0.050	*(0.084)*
Asia	0.974 *	*(0.119)*	0.408 °	*(0.227)*	0.648 *	*(0.126)*
Middle East	-0.288 °	*(0.159)*	-0.224	*(0.250)*	-0.177	*(0.134)*
Northern Africa			-0.099	*(0.338)*	-0.055	*(0.329)*
Southern Africa			-1.203 *	*(0.209)*	-1.160 *	*(0.128)*
Students (unit of observation)	266545		180544		447089	
Countries	39		38		77	
R^2	0.309		0.430		0.363	
			Science			
Central exams	0.343 *	*(0.112)*	0.471 *	*(0.174)*	0.417 *	*(0.108)*
North America	0.039	*(0.099)*	-0.075	*(0.158)*	0.023	*(0.108)*
South America	-0.509 *	*(0.129)*	-0.078	*(0.264)*	-0.199	*(0.120)*
Eastern Europe	0.084	*(0.118)*	0.114	*(0.167)*	0.130	*(0.092)*
Oceania	-0.092	*(0.113)*	-0.113	*(0.086)*	-0.067	*(0.080)*
Asia	0.127	*(0.149)*	0.098	*(0.170)*	0.063	*(0.125)*
Middle East	-0.390 *	*(0.126)*	-0.278	*(0.213)*	-0.317 *	*(0.103)*
Northern Africa			-0.639 °	*(0.330)*	-0.771 $^{+}$	*(0.326)*
Southern Africa			-1.715 *	*(0.163)*	-1.923 *	*(0.110)*
Students (unit of observation)	266545		180544		447089	
Countries	39		38		77	
R^2	0.219		0.407		0.317	

[a]Each column reports results from one regression. Dependent variable: TIMSS international math/science test score. All regressions control for all the family, resource, and institutional control variables reported in Table A1a. Clustering-robust standard errors in parentheses. All standard errors reported in this table take countries as the level of clustering. Significance levels (based on clustering-robust standard errors): * 1 percent, $^{+}$ 5 percent, ° 10 percent.

base estimates of Tables 3a and 3b as reasonably accurate estimates of the actual impact of central exams on student performance is weak.

5 How Do Central Exams Change the Working of the Education System?

The previous section presented results on the overall impact of central-exam systems on student performance without distinguishing between different impact channels. This section asks whether and how central exams exert their effects through several impact channels by changing the behavior of the different stakeholders in the education process. This is tested by analyzing whether different institutional features of the school system that relate to the influence of schools, teachers, and parents in the education process have different effects on student performance in systems with and without central exams. This evidence provides some indication on whether the behavior of the specific stake-

holders is affected by central-exam systems. If, for example, school autonomy in a specific decision-making area affects student performance negatively when no central exams are in place but positively under a central-exam system, this would suggest that the behavior of schools in this decision-making area is fundamentally altered by the existence of central exams. The first impact channel analyzed involves the way in which local autonomy of schools and teachers in several decision-making areas affects student performance (Section 5.1). Section 5.2 looks at whether the work of teachers and students gets focused on educational achievement through the implementation of central-exam systems. Section 5.3 uses a direct measure of parental influence to see how the impact of parental involvement differs between systems with and without central exams.

5.1 The Impact of School and Teacher Autonomy With and Without Central Exams

The question whether several institutional features have a different effect on student performance in systems with and without central exams is addressed by including interaction terms between these institutional features and central exams. Thus, the equations estimated in this section take the form

$$(2) \qquad T_{ilsc} = \alpha E_c + I_{lcs}\beta_1 + (E_c I_{lcs})\beta_2 + B_{ilcs}\beta_3 + R_{lcs}\beta_4 + \eta_c + \nu_{sc} + \varepsilon_{ilsc}$$

where the only change relative to equation (1) is the inclusion of interaction terms ($E_c I_{lsc}$) between central exams and the different institutional variables as additional explanatory variables.

Before analyzing the complete set of interaction terms between central exams and other institutional features, Figure 2 depicts selected examples of the interaction between central exams and local autonomy in several areas of decision-making corresponding to different cells in Table 1 (see Section 2.2). All estimates in Figure 2 are based on regressions using the math dataset that pools the TIMSS-95 and TIMSS-Repeat tests.[24] Each of the four pictures reports the performance of students in four situations: Students in systems without central exams whose school or teacher does not have autonomy in the specific decision-making area depicted by the picture; students without central exams but with local autonomy; students with central exams but without local autonomy; and students with central exams and with local autonomy. The estimates are presented in percent of a standard deviation in test scores, and the lowest-performing of the four categories in each picture has been set to zero.[25]

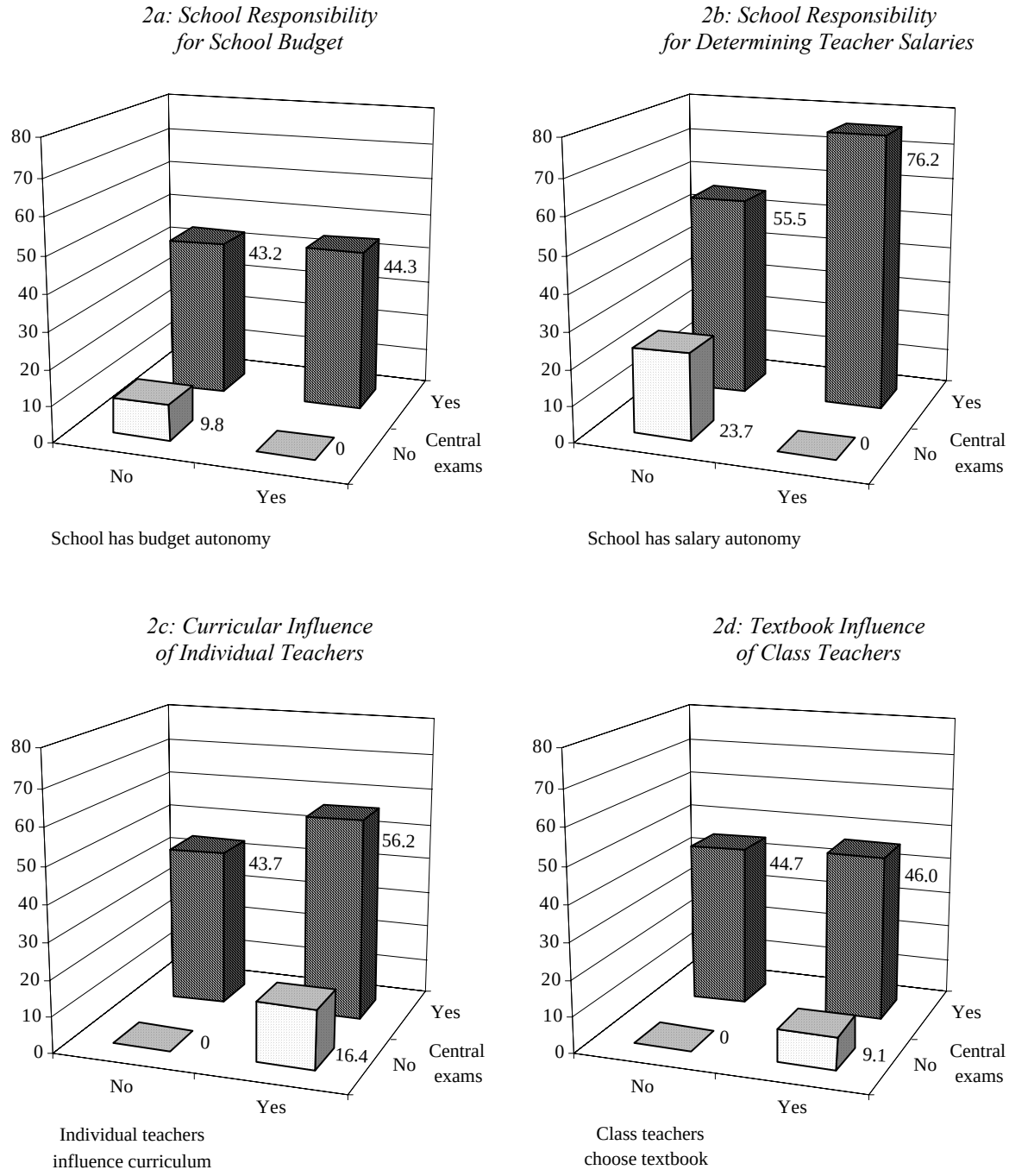
The first decision-making area analyzed is whether schools have autonomy over their budgets. This measure is based on a background-questionnaire item answered by the heads of schools who report whether formulating the school budget is primarily a school responsibility in their specific schools. Arguably, this is a case corresponding to cell [O1] of Table 1: The scope for opportunistic behavior on part of the school seems substantial in budgetary questions, as schools would seem to have other interests than purely furthering student performance when it comes to the money available to them. Furthermore, the scope for better-informed decision-making at the school level relative to some external level might be small in budgetary matters, as external agencies may even have superior knowledge in this area. Thus, one might expect that giving schools autonomy over formulating their own budget is detrimental to student performance when there is no system of central exams in place that helps in holding schools accountable for their decisions. However, once central-exam systems are in place to hold schools accountable, giving them budgetary autonomy might not lead to the detrimental opportunistic behavior.

---

[24] While the regressions on which the pictures in Figure 2 are based control for the whole set of family, resource, and institutional controls listed in Table A1a, they do not control for interaction terms between central exams and other variables. As will become evident in Table 8 below, the estimate of the genuine effect of central exams gets very imprecise once a whole set of interaction terms is introduced. Excluding other interaction terms allows to base the size of the bars depicting the impact in central-exam systems on reasonably exact, statistically significant estimates of the general effect of central exams.

[25] All estimates reported in Figure 2 are statistically significantly different from zero.

Figure 2: Central Exams and the Effects of School and Teacher Autonomy on Student Performance



Scale: TIMSS math performance relative to the lowest-performing category.

The results depicted in Figure 2a support this reasoning. In school systems without central exams, students performed 9.8 percent of a standard deviation better when their school did not have autonomy over the budget, suggesting that budgetary autonomy enables opportunistic behavior of schools when no central exams are in place. Students in schools with budgetary autonomy in central-exam systems performed 43.2 units better than students in a situation with school autonomy and without central exams, or 33.4 units better than students without school autonomy and without central exams.

Notably, there is no significant difference in student performance between schools with and without budgetary autonomy once a central-exam system is in place. This may suggest that central exams curb the opportunistic behavior of schools, and that there is no difference in how informed budgetary decisions are between school-based or external decision-makers. Alternatively, it may be the case that the negative impact of whatever opportunism is left in spite of the improved monitoring due to central exams is almost perfectly offset by any potential positive impact due to superior local knowledge. In either case, the detrimental effect of school autonomy in budgetary matters that exists in school systems without central exams is not existent in central-exam systems. This suggests that schools indeed respond to the altered incentive environment by behaving more favorably to student performance.

While school autonomy makes no difference to student performance in situations where opportunism is curbed and local knowledge is not important, it should have positive effects on student performance if opportunistic behavior is checked in and local knowledge is important to the task in question (cell [O2] in Table 1). This seems to be the case in the task of determining teacher salaries (Figure 2b). In systems without central exams, students in schools that have autonomy in determining teachers' salaries perform worse than students in schools that do not have salary autonomy. This might reflect that schools again behave opportunistically in this decision-making area where money is involved, as long as they cannot be held accountable to their behavior. In systems with central exams, by contrast, students in schools with salary autonomy perform better, not worse, than students in schools without salary autonomy. That is, the effect of school autonomy is reversed once central exams are in place. It seems that in salary decisions, heads of school know better than any external agency which teacher worked hard and deserves a bonus or pay rise and which teacher does not. Again, the evidence on salary autonomy strongly suggests that schools seem to change their behavior once central exams enable monitoring of educational outcomes.

The last two pictures of Figure 2 deal with the influence of individual teachers in decision-making areas where local knowledge seems to be important, but where the scope for opportunistic behavior seems limited. The evidence presented in Figure 2c is based on a background-questionnaire item answered by the heads of school on how much influence each teacher individually (as opposed to teachers collectively and to other educational stakeholders) has in determining the curriculum that is taught in their schools. The picture contrasts schools where individual teachers had a lot of influence on the curriculum to schools where teachers had no, little, or only some curricular influence. Both in systems with and without central exams, students in schools where individual teachers had a lot of influence on the curriculum scored significantly better than students in schools where they did not have a lot of influence. The difference between systems with and without central exams in the advantage of schools with teacher influence is not statistically significant. This suggests that curricular influence of individual teachers is an example of cell [N2a] in Table 1: There does not seem to be much scope for opportunistic behavior on part of individual teachers in this area, individual teachers' knowledge on how to teach the curriculum seems to be substantial, and central exams do not seem to limit the positive impact of teacher autonomy. However, performance in central-exam systems is still substantially superior to performance in systems without central exams, a differential impact that presumably works through other channels.

Figure 2d presents evidence on teacher autonomy in the choice of textbooks, based on a background-questionnaire item answered by the math teachers on how much influence they have on the specific textbook to be used. Students whose teacher reported a lot of influence on textbook choice scored better than students of teachers without a lot of textbook influence in systems without central exams. By contrast, in systems with central exams there was no statistically significant difference between teachers with and without autonomy in the choice of textbooks. This may reflect the situation of cell [N2b] in Table 1, where there is not much scope for opportunism—the choice of a poor textbook would probably hurt the teachers themselves as much as the students—, where the local

knowledge of teachers is important on which textbook might be best for their students, and where central exams to some extent limit teachers' capabilities to make the best choices.

It should be borne in mind, however, that the most obvious pattern in all pictures of Figure 2 is that student performance is substantially better when central exams are in place. The change in school and teacher behavior reflected in the different impact of school and teacher autonomy between systems with and without central exams seems to be one of several channels through which this superior performance comes about. Furthermore, the positive impact of central exams is especially apparent in decisions where opportunistic behavior can be curbed, and this is especially the case wherever financial resources are involved, such as budgetary and salary decisions.

Tables 8a and 8b present evidence on including a complete set of interaction terms between central exams and other institutional features of the school system as in equation (2) for the pooled TIMSS-95/TIMSS-Repeat data in math and in science. The first column in both tables reports the coefficient estimates β_1 on the different institutions I, and the second column reports the estimates β_2 on the interaction term EI between each institution and central exams of the same regression. The last two columns report equivalent evidence for a specification that additionally controls for interaction terms between student characteristics and central exams.

The pattern of results presented in Figure 2 is robust against the inclusion of other institutional interactions and of family-background interactions, and that the pattern in science is very similar to the pattern in math. (Note that to determine the combined impact of central exams and an institutional characteristic, the three coefficient estimates on central exams, on the institutional coefficient, and on their interaction term have to be added.) In the richest specifications, the point estimate of the coefficient on central exams, which reflects the effect of central exams in the absence of all the characteristics depicted by the institutional and family-background variables, is no longer statistically significant as the standard error increases.

In addition to the interaction effects discussed in Figure 2, school autonomy in purchasing supplies has a positive effect on student performance that is somewhat smaller in central-exam systems than in systems without central exams, reflecting cell [N2b] in Table 1. The pattern for school autonomy in hiring teachers is less clear, with the effect in math being positive in systems without central exams but about zero with central exams, and an opposite finding in science. Teacher autonomy over money for supplies has a negative impact on student performance when no central-exam system is in place but a positive impact with central exams, reflecting the case of opportunism and important local knowledge of cell [O2] in Table 1—teachers' influence on the money for supplies seems to get well channeled once central exams introduce accountability. Teacher autonomy in the choice of the subject matter to be covered in class has a negative impact on student performance that is substantially lowered in math when central exams are in place, suggesting that there is large scope for opportunism on part of the teachers to determine their own work-load in this decision-making area (cell [O1] of Table 1 for math).

As argued in Section 2.2, teachers' influence may be especially prone to opportunistic behavior when exerted by teachers as an interest group. Accordingly, in systems without central exams, teachers' influence on the curriculum is detrimental to student performance once it is exerted by teachers of the same subject as a group, by teachers collectively for the school, or by teacher unions. In the case of teachers grouping together within a school (teachers of the same subject and all school teachers collectively), this negative effect is substantially mitigated when central exams are in place, reflecting a situation comparable to cell [O1] of Table 1. The negative influence of teachers acting as unions to influence the curriculum, however, is even more detrimental in systems with central exams than in systems without central exams. This suggests that central-exam systems are especially susceptible to the group interests of teachers once these are pursued at the system level, as their interests might then water down the design and implementation of the central-exam systems themselves.

Table 8a: Interaction Effects of Central Exams with Other Institutional Settings (Math)[a]

	With institutional interactions		With instit. and student interactions	
	Coefficient	Interaction with c.e.	Coefficient	Interaction with c.e.
Central exams (c.e.)	0.390 ° *(0.197)*[b]		0.664 *(1.041)*[b]	
Institutional settings				
School responsibility				
School budget	-0.071 + *(0.028)*	0.080 + *(0.035)*	-0.069 + *(0.028)*	0.077 + *(0.035)*
Purchasing supplies	0.070 + *(0.033)*	-0.057 *(0.050)*	0.071 + *(0.032)*	-0.057 *(0.050)*
Hiring teachers	0.218 * *(0.027)*	-0.207 * *(0.033)*	0.216 * *(0.026)*	-0.202 * *(0.031)*
Determining teacher salaries	-0.279 * *(0.037)*	0.497 * *(0.042)*	-0.283 * *(0.036)*	0.502 * *(0.041)*
Teachers' influence				
Class teacher has strong influence on				
Money for supplies	-0.260 * *(0.053)*	0.304 * *(0.065)*	-0.247 * *(0.051)*	0.291 * *(0.063)*
Kind of supplies	0.033 *(0.029)*	-0.040 *(0.038)*	0.030 *(0.028)*	-0.035 *(0.038)*
Subject matter	-0.120 * *(0.024)*	0.085 * *(0.028)*	-0.123 * *(0.023)*	0.087 * *(0.028)*
Textbook	0.116 * *(0.032)*	-0.117 * *(0.037)*	0.116 * *(0.031)*	-0.117 * *(0.036)*
Strong influence on curriculum				
Teacher individually	0.161 * *(0.021)*	-0.057 + *(0.028)*	0.146 * *(0.021)*	-0.039 *(0.027)*
Subject teachers	-0.054 + *(0.025)*	0.034 *(0.032)*	-0.050 + *(0.024)*	0.028 *(0.031)*
School teachers collectively	-0.158 * *(0.021)*	0.073 * *(0.028)*	-0.147 * *(0.021)*	0.065 + *(0.028)*
Teacher unions	-0.063 *(0.053)*	-0.319 * *(0.086)*	-0.085 *(0.054)*	-0.295 * *(0.087)*
Students' incentives				
Scrutiny of testing	0.037 * *(0.006)*	-0.013 ° *(0.007)*	0.037 * *(0.006)*	-0.012 ° *(0.007)*
Homework	0.012 *(0.007)*	0.017 + *(0.009)*	0.014 + *(0.007)*	0.015 ° *(0.009)*
Parents' influence				
Uninterested parents limit teaching	-0.098 * *(0.035)*	-0.075 ° *(0.042)*	-0.099 * *(0.033)*	-0.077 ° *(0.041)*
Interested parents limit teaching	-0.198 * *(0.054)*	0.201 * *(0.061)*	-0.177 * *(0.051)*	0.178 * *(0.059)*
Student and family characteristics				
Upper grade	0.372 * *(0.010)*		0.255 * *(0.019)*	0.175 * *(0.023)*
Above upper grade	0.988 * *(0.036)*		0.949 * *(0.037)*	
Age	-0.129 * *(0.007)*		-0.122 * *(0.012)*	-0.008 *(0.014)*
Sex	-0.074 * *(0.006)*		-0.063 * *(0.013)*	-0.015 *(0.015)*
Born in country	0.174 * *(0.013)*		0.203 * *(0.021)*	-0.042 *(0.029)*
Living with both parents	0.101 * *(0.007)*		0.052 * *(0.011)*	0.074 * *(0.014)*
Parent born in country	-0.014 *(0.012)*		0.148 * *(0.021)*	-0.216 * *(0.028)*
Parents' education				
Finished primary	0.136 * *(0.009)*		0.226 * *(0.017)*	-0.142 * *(0.021)*
Secondary	0.203 * *(0.010)*		0.319 * *(0.019)*	-0.179 * *(0.023)*
Finished university	0.391 * *(0.011)*		0.490 * *(0.020)*	-0.158 * *(0.026)*
Books at home				
11-25	0.145 * *(0.009)*		0.155 * *(0.021)*	-0.009 *(0.024)*
26-100	0.403 * *(0.010)*		0.356 * *(0.023)*	0.068 * *(0.026)*
101-200	0.527 * *(0.010)*		0.499 * *(0.024)*	0.042 *(0.027)*
More than 200	0.596 * *(0.011)*		0.567 * *(0.025)*	0.044 *(0.028)*
Further controls				
Centralization [2 variables]		✓		✓
Community location, GDP [3 var.]		✓		✓
Resources/teachers [13 variables]		✓		✓
Students (unit of observation)	447089		447089	
Schools (PSUs)	12175		12175	
Countries	77		77	
R^2	0.294		0.296	

[a]Every two columns headed "Coefficient" and "Interaction with c.e." together report the results of one regression. "Coefficient" reports the coefficient on the variable labeled in each row, while "Interaction with c.e." reports the coefficient on the interaction term between central exams and the variable labeled in the row. Dependent variable: TIMSS international math test score. Clustering-robust standard errors in parentheses (schools as level of clustering unless noted otherwise). Significance levels (based on clustering-robust standard errors): * 1 percent, + 5 percent, ° 10 percent. – [b]Standard error has countries as the level of clustering.

Table 8b: Interaction Effects of Central Exams with Other Institutional Settings (Science)[a]

	With institutional interactions		With institutional and student interactions	
	Coefficient	Interaction with c.e.	Coefficient	Interaction with c.e.
Central exams (c.e.)	0.042 *(0.209)*[b]		0.981 *(1.004)*[b]	
Institutional settings				
School responsibility				
School budget	-0.121 * *(0.026)*	0.163 * *(0.034)*	-0.120 * *(0.026)*	0.161 * *(0.035)*
Purchasing supplies	0.165 * *(0.030)*	-0.072 *(0.054)*	0.156 * *(0.031)*	-0.062 *(0.054)*
Hiring teachers	-0.013 *(0.019)*	0.064 + *(0.025)*	0.003 *(0.019)*	0.046 ° *(0.026)*
Determining teacher salaries	-0.073 * *(0.026)*	0.280 * *(0.031)*	-0.082 * *(0.026)*	0.292 * *(0.031)*
Teachers' influence				
Class teacher has strong influence on				
Money for supplies	-0.062 ° *(0.036)*	0.129 * *(0.045)*	-0.069 ° *(0.036)*	0.136 * *(0.045)*
Kind of supplies	0.054 * *(0.020)*	-0.030 *(0.029)*	0.060 * *(0.020)*	-0.037 *(0.029)*
Subject matter	-0.041 + *(0.017)*	-0.013 *(0.022)*	-0.046 * *(0.017)*	-0.007 *(0.022)*
Textbook	0.061 * *(0.018)*	-0.096 * *(0.026)*	0.063 * *(0.018)*	-0.099 * *(0.026)*
Strong influence on curriculum				
Teacher individually	0.150 * *(0.018)*	-0.082 * *(0.025)*	0.145 * *(0.018)*	-0.074 * *(0.025)*
Subject teachers	-0.056 * *(0.021)*	0.083 * *(0.028)*	-0.058 * *(0.021)*	0.082 * *(0.028)*
School teachers collectively	-0.160 * *(0.019)*	0.152 * *(0.026)*	-0.153 * *(0.019)*	0.144 * *(0.026)*
Teacher unions	-0.040 *(0.052)*	-0.349 * *(0.093)*	-0.067 *(0.051)*	-0.300 * *(0.091)*
Students' incentives				
Scrutiny of testing	-0.008 ° *(0.005)*	0.017 * *(0.006)*	-0.008 ° *(0.005)*	0.017 * *(0.006)*
Homework	-0.046 * *(0.010)*	0.062 * *(0.015)*	-0.043 * *(0.010)*	0.060 * *(0.015)*
Parents' influence				
Uninterested parents limit teaching	-0.017 *(0.028)*	-0.177 * *(0.039)*	-0.031 *(0.028)*	-0.160 * *(0.039)*
Interested parents limit teaching	-0.102 + *(0.041)*	0.171 * *(0.052)*	-0.098 + *(0.041)*	0.170 * *(0.052)*
Student and family characteristics				
Upper grade	0.440 * *(0.009)*		0.370 * *(0.012)*	0.122 * *(0.017)*
Above upper grade	0.991 * *(0.031)*		0.954 * *(0.031)*	
Age	-0.107 * *(0.006)*		-0.073 * *(0.006)*	-0.064 * *(0.010)*
Sex	-0.159 * *(0.005)*		-0.121 * *(0.009)*	-0.058 * *(0.012)*
Born in country	0.210 * *(0.013)*		0.281 * *(0.021)*	-0.115 * *(0.030)*
Living with both parents	0.073 * *(0.007)*		0.026 * *(0.010)*	0.079 * *(0.014)*
Parent born in country	0.047 * *(0.012)*		0.114 * *(0.023)*	-0.086 * *(0.029)*
Parents' education				
Finished primary	0.074 * *(0.008)*		0.123 * *(0.015)*	-0.073 * *(0.019)*
Secondary	0.166 * *(0.009)*		0.193 * *(0.015)*	-0.047 + *(0.020)*
Finished university	0.354 * *(0.010)*		0.385 * *(0.018)*	-0.055 + *(0.023)*
Books at home				
11-25	0.164 * *(0.009)*		0.152 * *(0.016)*	0.019 *(0.020)*
26-100	0.414 * *(0.010)*		0.361 * *(0.017)*	0.080 * *(0.022)*
101-200	0.568 * *(0.011)*		0.490 * *(0.018)*	0.118 * *(0.023)*
More than 200	0.647 * *(0.011)*		0.559 * *(0.019)*	0.135 * *(0.024)*
Further controls				
Centralization [2 variables]		✓		✓
Community location, GDP [3 var.]		✓		✓
Resources/teachers [13 variables]		✓		✓
Students (unit of observation)	447089		447089	
Schools (PSUs)	12175		12175	
Countries	77		77	
R^2	0.264		0.266	

[a]Every two columns headed "Coefficient" and "Interaction with c.e." together report the results of one regression. "Coefficient" reports the coefficient on the variable labeled in each row, while "Interaction with c.e." reports the coefficient on the interaction term between central exams and the variable labeled in the row. Dependent variable: TIMSS international science test score. Clustering-robust standard errors in parentheses (schools as level of clustering unless noted otherwise). Significance levels (based on clustering-robust standard errors): * 1 percent, + 5 percent, ° 10 percent. – [b]Standard error has countries as the level of clustering.

5.2 The Impact of Regular Testing and Homework With and Without Central Exams

Teachers often use devices to monitor students' efforts in order to increase their performance. Two such devices are regular testing of students' educational progress and the assignment of homework to have students practice their knowledge. In central-exam systems, the impact of such devices on student performance might be altered in two ways: First, teachers' incentives are aligned with student performance due to their own increased monitoring by parents and heads of schools, which should increase teachers' efforts to focus these devices on ensuring high student performance. Second, as students themselves get better monitored, their own effort should increase and get better focused on educational achievement (see Section 2.2 above).

Scrutiny of testing is measured discretely by teachers' responses on how many hours per week they normally spend outside the school day preparing or grading student tests or exams. Similarly, homework assignment is measured discretely in hours per week based on teachers' reports on how often and for how many minutes they usually assign homework. In math, both scrutiny of testing and homework have positive effects on student performance both in systems with and without central exams (Table 8a). The effect of testing is slightly smaller with central-exam systems, which might reflect that teacher testing comes in addition to central-exam testing in central-exam systems while it is the only way of testing in systems without central exams. By contrast, the positive effect of homework assignment is doubled in central-exam systems. In science, both monitoring devices actually have a slightly negative effect on student performance in systems without central exams. The effect is turned into a positive one once central exams are in place (Table 8b).

This shows that monitoring devices such as regular testing and homework assignment do not seem to further student performance strongly as long as agents' incentives are not aligned with the goal of increased student performance. As long as this is not the case, the design and content of these devices do not seem to be well focused, a problem that is especially severe in the case of subjects whose content may be less coherent in the absence of explicit standards (for example, science as compared with math). Given the alignment of incentives with student performance in central-exam systems, teachers' and students' efforts in the design of and performance on tests and homework seem to get better focused on enhancing students' educational achievement.

5.3 The Impact of Parental Influence With and Without Central Exams

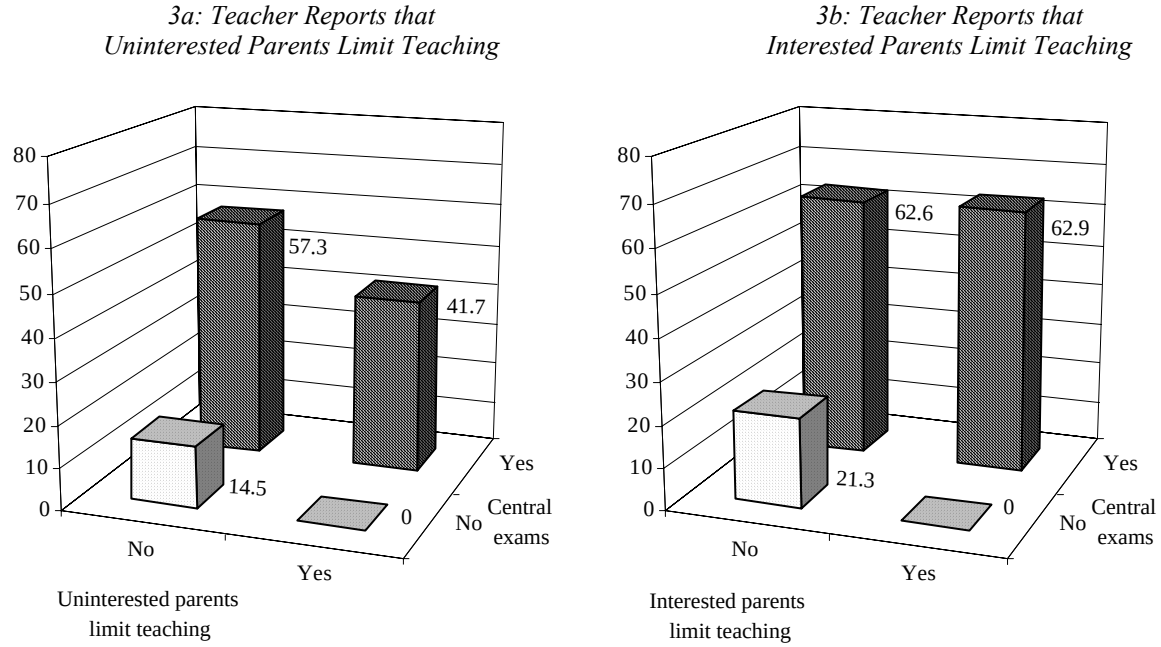
All effects discussed so far may be linked to changes in the behavior of parents who are able to increase the monitoring of educational achievement once they have the information generated by central exams. As argued in Section 2.2, this positive effect of central exams will be especially salient with parents who are strongly concerned with their child's educational progress, but not as much with parents who are less concerned about their child's education. Two measures contained in the TIMSS teacher background questionnaires may help to shed some light on this differential impact. First, teachers reported to what extent, in their view, parents *uninterested* in their child's learning and progress limit how the teachers teach their class. Second, teachers also reported whether their teaching is limited by parents *interested* in their child's progress.

The math performance of students in the different situations is depicted in Figure 3.[26] Students whose teachers reported that their teaching was not substantially limited by uninterested parents performed better than students whose teachers reported that their teaching was limited by uninterested

[26]As was the case in Figure 2, the regressions on which Figure 3 are based control for family, resource, and institutional variables, but not for interaction terms between central exams and other variables.

parents, irrespective of whether a central-exam system was in place (Figure 3a).[27] The results are different for the involvement of parents who are interested in their child's progress, however. In systems without central exams, students whose teachers reported that their teaching was limited a lot by interested parents again performed worse. But in central-exam systems, students whose teachers reported that interested parents limited how they teach their class performed just as well as students whose teachers did not say so. That is, even though teachers judged the intrusion of interested parents as limiting their teaching, student performance in fact did not suffer from this "limitation."

Figure 3: Central Exams and the Involvement of Parents: Their Effects on Student Performance



Scale: TIMSS math performance relative to the lowest-performing category.

In science, the negative impact of uninterested parents was even more negative in systems with central exams than in systems without central exams (Table 8b). For interested parents limiting teaching, the negative effect in systems without central exams is turned around to be positive when central exams are in place. Even though teachers complained that their teaching was limited by the involvement of interested parents, the performance of students was actually furthered by this parental intervention.

While the involvement of interested parents may limit student performance in systems without central exams because parents do not have well-founded information on which to base their interventions, central-exam systems seem to ensure that interested parents have the information necessary to intervene properly. Parents uninterested in their child's educational progress do not seem to make use of this information, and their lack of interest hurts students' educational performance. But it seems that the involvement of interested parents can never go all the way to being detrimental when central exams are in place, even when teachers might judge it to be so. While there is no data to estimate the effect of the involvement of interested parents when it is approved by the teachers, it seems likely that this would be even superior for teaching and learning.

---

[27] In the specification of the estimation equation that controls for all other institutional interaction effects (Table 8a), the negative impact of uninterested parents with central exams is even worse than without central exams.

6 Conclusion: Do Central Exams Lead to Real Gains in Knowledge?

The international evidence based on TIMSS-95 and TIMSS-Repeat confirms that central exams are a powerful accountability device. Student performance in math and science is substantially higher in school systems with central exams than without central exams, and this is true for students from all performance quartiles and family backgrounds. Parents, administrators, schools, teachers, and students all appear to respond to the changed incentive environment created by central exams by behaving more favorably to students' educational achievement. Parental involvement becomes more informed and effective. Opportunistic behavior of schools and teachers is curbed, so that local autonomy in many decision-making areas becomes an attractive feature of a school system. And the efforts of teachers and students are more concentrated on the goals of the education system as represented in the exam content.

When considering what individual countries can specifically learn from this evidence, the specific features of the school systems and the pressing policy questions have to be borne in mind. In Germany, with its rather bureaucratic school system, the evidence suggests in particular that before rushing into overall decentralization, German states (Länder) which do not have a central-exam system should consider implementing one in order to ensure that school autonomy in other decision-making areas produces beneficial effects. Without central exams, giving schools autonomy in decision-making areas like budgetary and salary decisions and the choice of the covered subject matters could backfire. Central exams seem to be a prerequisite for a decentralized school system to function properly. The fact that central exams particularly improve the performance of students from weak parental backgrounds suggests that concerns about the distribution of student performance in Germany could to some extent be mitigated by a more general introduction of central exams.

In the case of the United States, it is especially relevant to analyze how central exams work in systems with a high level of local autonomy, as the U.S. school system is to a large extent locally controlled and funded and has no general centralized system of wage bargaining, contracting, or teacher assignment. The results suggest that central exams are especially capable of bringing out the positive aspects of local autonomy, while mitigating its negative consequences. In some cases, central exams also seem to limit the ability of local decision-makers to make appropriate decisions. However, such limitations are far outweighed by their positive incentive effects.

One criticism often given to all test-based accountability systems is that they might lead to "teaching to the test" rather than real increases in students' knowledge. As this is obviously an important issue, three comments on this question are in order. First, the performance information used in this paper does not originate from the accountability-creating test. Instead, the measures of student performance in math and science are students' test scores in the international TIMSS tests, which were accepted by representatives of all participating countries as covering the basic math and science curriculum for middle-school students. Even more importantly, no stakes for students or schools were attached to the TIMSS tests. If teachers were just teaching how to take the specific central exam, and if students were just learning how to take this specific exam, then this should not affect student performance on the TIMSS tests. Thus, the fact that students in countries with central-exam systems did perform substantially better on the TIMSS tests allows the inference that the central exams indeed caused superior math and science knowledge of the students and not just an increased capability of taking the one specific central exam.

Second, the valuation of "teaching to the test" depends crucially on what exactly is meant by this (cf. Hoxby 2002a). If, as in the previous paragraph, it refers to just teaching how to take a specific test ("teaching *the* test")—for example, by giving students answers to specific questions that will probably be asked in the test—as opposed to increasing students' knowledge in the subjects, this is clearly not an outcome to be aimed for. If, by contrast, it refers to teaching being more focused on the content

areas covered by the test ("teaching *towards* the test") as opposed to teaching other content areas that are not part of the test, this is precisely consistent with the aims of implementing a central-exam system: Central exams are meant to focus attention on the goals of the education system, and as long as these goals are clearly spelled out and as the central exams cover exactly these content areas, this helps in aligning the working of the school system with its goals.

Third, much of the capacity of central exams to lead to real knowledge gains depends on the quality of the exam. It is possible to devise the exams in a way that makes teaching how to take the specific exam hardly feasible. Having the exam performed by outside proctors and using fresh questions each year will assure that "teaching *the* test" is not possible (Hoxby 2002a). Furthermore, central testing by no means requires that the test be all multiple choice. Many countries have central-exam systems requiring much individual creativity. It is also possible to combine central testing with the freedom of students to choose among subject areas, while at the same time maintaining the pivotal incentive mechanism created by external testing. These questions of test quality are certainly beyond the scope of this paper. It should be borne in mind, however, that they are not fundamentally different between central exams and any other examination system.

As a final assessment, the relative merits of central exams as an accountability device may be compared to other accountability systems, such as teacher merit pay, school-based accountability systems, or district report cards. There is much discussion in the literature about which educational stakeholders should be targeted by accountability systems. Much of the current U.S. discussion on educational accountability seems to favor rewards for high-achieving schools and/or sanctions for failing schools. For example, Ladd (2001: 386) argues that "subject to some important qualifications related to funding and capacity, schools are an appropriate unit for accountability purposes and have clear advantages compared to other possible units of accountability, such as school districts, individual teachers, and students." In contrast to this recommendation, central-exam systems primarily target the individual students who take the central exam (cf. Hanushek 2002). However, the arguments and evidence presented in this paper show that the incentives created by central-exam systems extend far beyond the individual student. With central exams providing the information necessary to monitor educational outcomes, all stakeholders are more likely to face consequences for their behavior. Thus, central exams not only have the direct effect of changing students' incentives, but they also work indirectly to change incentives all the way up the agency "ladder" spanning from students over teachers and schools to administrators. As all these stakeholders respond to incentives, their behavior becomes more closely aligned with furthering students' educational performance.

The practical merits of other accountability systems are less clear. Performance-related pay for teachers has generally been deemed a failure in the American public school system (cf. Murnane and Cohen 1986; Ballou 2001). Several recent studies have hinted at substantial implementation problems facing school-based accountability systems that rely on value-added measures of performance. For example, value-added measures of a school's performance at a particular grade have been shown to vary substantially in ways unrelated to school performance, both due to ability differences in the student sample and due to one-time factors (Kane and Staiger 2001; Figlio and Page 2002). Additionally, Ladd and Walsh (2002) find that even the more sophisticated value-added measures of school effectiveness currently implemented, which follow the performance of students from year to year, fail to thoroughly account for resource differences and measurement error in the test-score data. Since measurement errors are amplified when the data used is based on changes rather than levels, this problem is especially severe for value-added measures. However, one would not want to base schools' performance assessments on level measures of their students' performance, which are strongly determined by the students' social background. Thus, both school-based accountability systems based on value-added measures of performance and those based on level measures of performance could lead to distorted incentives and arbitrary performance evaluations for schools. By contrast, student-based central-exam systems, which are based on level measures of performance, are less prone to

arbitrariness and create incentives that induce each student to get the best possible performance out of his or her ability and social background.

Despite their apparent connotation of centralizing decision-making, central-exam systems ironically may require less central regulation and allow more flexibility at the local level. For external-exam systems to exert their beneficial incentive effects, it is not required that any central person or agency has detailed knowledge of the educational production process in every school. Central administrators may in practice lack the necessary information to intervene in a beneficial way—and the solutions for different failing schools may in fact differ depending on backgrounds, customs, and local experiences. Rather than trying to micro-manage schools by central regulators, external exams change the system so that the incentives of all stakeholders are better aligned with the goals of the system. If adequately motivated to improve performance and equipped with valid performance information, local stakeholders may actually be better equipped than any central agency to evaluate accountability and thus to reward or punish performance. Given the implementation problems of accountability systems that rely on central regulation, evaluation, and intervention, the relative merits of external-exam systems as an accountability device make them a highly attractive policy.

Appendix: Construction of the TIMSS-Repeat Database

The TIMSS-Repeat database used in this paper was constructed in a similar way to Wößmann's (2002a) TIMSS-95 database. The database construction starts by combining data from the TIMSS-Repeat math and science performance files with data from the TIMSS-Repeat student, teacher, and school background-questionnaire files for all participating countries. If a student had more than one teacher in math or science, he or she was assigned the teacher who instructed him or her for the longest period of time.

While complete performance data was available for all students, various variables from the different background questionnaires contain missing values. I decided to exclude student observations with an excessive amount of missing data and to impute values for the remaining missing data. In order to determine the observations to be excluded, the availability of a set of core variables in each background questionnaire was observed, which were 10 variables in the student background questionnaire, 16 variables each in the math and science teacher questionnaire, and 25 variables in the school questionnaire. If in all four questionnaires more than half of the core variables were missing, the student was dropped entirely from the sample. This was the case for 156 students, scattered across seven countries. For the remaining 180,544 students, more than half of the core variables were answered in at least one of the questionnaires.

As one would give away a lot of valuable information and presumably introduce substantial sample-selection bias if one dropped also these students from the sample—because, for example, the teachers of a specific student might have answered their questionnaires poorly, but the student and school questionnaire of this student may be available and well-answered—I chose to impute values in these remaining cases of missing values. Using a set of 22 basic variables that were available for nearly all students as predictor variables,[28] an ordered probit model was estimated to forecast the probability of occurrence associated with the different categories of each qualitative survey variable, based on the

[28] These basic predictor variables were: Students' sex, age, whether the student was born in the country, four dummies on the number of books in the students' home, three dummies on the community location, three dummies on the status of availability of materials in the school, teachers' age, sex, and year of experience, four dummies for teachers' education, the gross national income of the country, and expenditure per student in the country. In the few cases where values on a predictor variable were missing, these were imputed through class, school, or country means, whichever was the lowest level with available data, before the imputation of the other variables.

observations with available values on this variable. For the observations with missing values on this variable, the category with the highest probability—based on the coefficients estimated by the ordered probit model and on the basic predictor variables of these observations—was imputed. Similarly, the category with the highest probability of occurrence based on a probit model was imputed for missing values of dichotomous variables, and missing values of discrete variables were imputed using a least-squares model.[29] In the now complete database that contains imputed values for missing data, the qualitative questionnaire data were transformed into dummy variables (indicating whether a specific state was given or not) for the subsequent estimations.

[29] See the appendix of Wößmann (2002a) for details on the imputation technique.

Table A1a: Complete Base Results (Math)[a]

	TIMSS-95		TIMSS-Repeat		Pooled	
	Coef.	*S.E.*	Coef.	*S.E.*	Coef.	*S.E.*
Central exams	0.409*	*(0.135)*[b]	0.470*	*(0.135)*[b]	0.427*	*(0.098)*[b]
Institutional settings						
School responsibility						
School budget	-0.071*	*(0.024)*	-0.008	*(0.021)*	-0.021	*(0.016)*
Purchasing supplies	-0.002	*(0.033)*	0.106*	*(0.031)*	0.029	*(0.025)*
Hiring teachers	0.072*	*(0.017)*	0.087*	*(0.017)*	0.064*	*(0.012)*
Determining teacher salaries	0.122*	*(0.020)*	0.041+	*(0.018)*	0.104*	*(0.014)*
Teachers' influence						
Strong influence on curriculum						
Teacher individually	0.128*	*(0.017)*	0.073*	*(0.019)*	0.137*	*(0.013)*
Subject teachers	-0.068*	*(0.019)*	0.014	*(0.020)*	-0.026°	*(0.014)*
School teachers collectively	-0.124*	*(0.018)*	-0.050*	*(0.019)*	-0.111*	*(0.013)*
Teacher unions	-0.246*	*(0.051)*	-0.327*	*(0.069)*	-0.306*	*(0.051)*
Class teacher has strong influence on						
Money for supplies	0.003	*(0.033)*	-0.032	*(0.039)*	-0.031	*(0.028)*
Kind of supplies	-0.036°	*(0.019)*	0.021	*(0.025)*	0.001	*(0.017)*
Subject matter	-0.016	*(0.015)*	-0.073*	*(0.017)*	-0.061*	*(0.012)*
Textbook	0.060*	*(0.019)*	0.024	*(0.019)*	0.036+	*(0.014)*
Students' incentives						
Scrutiny of testing (hours per week)	0.049*	*(0.004)*	0.012*	*(0.005)*	0.031*	*(0.003)*
Homework (hours per week)	0.005	*(0.005)*	0.038*	*(0.006)*	0.023*	*(0.004)*
Parents' influence						
Uninterested parents limit teaching	-0.091*	*(0.026)*	-0.172*	*(0.024)*	-0.153*	*(0.019)*
Interested parents limit teaching	-0.126*	*(0.037)*	0.035	*(0.031)*	-0.052+	*(0.025)*
Centralization						
Central curriculum	0.128	*(0.115)*[b]	0.211+	*(0.093)*[b]	0.163°	*(0.096)*[b]
Central textbook approval	0.121	*(0.112)*[b]	0.413*	*(0.112)*[b]	0.160°	*(0.082)*[b]
Student and family characteristics						
Upper grade	0.378*	*(0.010)*			0.368*	*(0.010)*
Above upper grade	0.981*	*(0.038)*			1.019*	*(0.036)*
Age (years)	-0.114*	*(0.007)*	-0.170*	*(0.007)*	-0.133*	*(0.006)*
Sex (female)	-0.073*	*(0.008)*	-0.071*	*(0.007)*	-0.073*	*(0.006)*
Born in country	0.083*	*(0.013)*	0.212*	*(0.020)*	0.176*	*(0.013)*
Living with both parents	0.109*	*(0.008)*	0.118*	*(0.010)*	0.107*	*(0.007)*
Parent born in country	0.018	*(0.016)*	-0.037°	*(0.019)*	-0.015	*(0.013)*
Parents' education						
Finished primary	0.249*	*(0.011)*	0.074*	*(0.016)*	0.144*	*(0.009)*
Secondary	0.146*	*(0.012)*	0.244*	*(0.018)*	0.204*	*(0.010)*
Finished university	0.397*	*(0.013)*	0.435*	*(0.019)*	0.393*	*(0.011)*
Books at home						
11-25	0.113*	*(0.014)*	0.141*	*(0.010)*	0.149*	*(0.009)*
26-100	0.355*	*(0.014)*	0.364*	*(0.011)*	0.406*	*(0.010)*
101-200	0.475*	*(0.016)*	0.498*	*(0.012)*	0.531*	*(0.010)*
More than 200	0.554*	*(0.016)*	0.576*	*(0.014)*	0.602*	*(0.011)*
Community location						
Geographically isolated area	-0.216*	*(0.034)*	0.051	*(0.040)*	-0.107*	*(0.029)*
Close to the center of a town	0.036+	*(0.015)*	0.124*	*(0.015)*	0.081*	*(0.011)*
GDP per capita (1000 intl. $)	0.037+	*(0.017)*[b]	0.083*	*(0.014)*[b]	0.055*	*(0.011)*[b]

(Table continued on next page)

Table A1a (continued)

	TIMSS-95		TIMSS-Repeat		Pooled	
	Coef.	*S.E.*	Coef.	*S.E.*	Coef.	*S.E.*
Resources and teachers						
Expenditure per student (1000 intl. $)	-0.033	*(0.054)*[b]	-0.185*	*(0.059)*[b]	-0.083+	*(0.041)*[b]
Class size (no. of students)	0.010*	*(0.001)*	-0.001	*(0.001)*	0.004*	*(0.001)*
Student-teacher ratio (10 students)	0.002+	*(0.001)*	-0.005*	*(0.001)*	-0.003*	*(0.001)*
No shortage of materials	0.091*	*(0.016)*	0.093*	*(0.018)*	0.085*	*(0.012)*
Great shortage of materials	-0.029	*(0.023)*	0.002	*(0.021)*	-0.033+	*(0.017)*
Instruction time (100 hours per year)	0.018*	*(0.005)*			0.020*	*(0.005)*
Instruction time (hours per week)			-0.005+	*(0.002)*	-0.007*	*(0.001)*
Teacher characteristics						
Teacher's sex (female)	0.059*	*(0.013)*	0.068*	*(0.016)*	0.092*	*(0.011)*
Teacher's age (years)	-0.006*	*(0.001)*	-0.002	*(0.001)*	-0.005*	*(0.001)*
Teacher's experience (years)	0.009*	*(0.001)*	0.008*	*(0.001)*	0.010*	*(0.001)*
Teacher's education						
Secondary only	0.163*	*(0.054)*	0.216	*(0.238)*	0.234*	*(0.055)*
BA or equivalent	0.208*	*(0.052)*	0.510+	*(0.235)*	0.376*	*(0.053)*
MA/PhD	0.299*	*(0.055)*	0.611*	*(0.236)*	0.468*	*(0.055)*
Other post-secondary			0.439°	*(0.238)*	0.333*	*(0.064)*
Constant	4.133*	*(0.129)*	3.830*	*(0.273)*	3.751*	*(0.113)*
Students (unit of observation)	266545		180544		447089	
Schools (PSUs)	6107		6068		12175	
Countries	39		38		77	
R^2	0.238		0.362		0.285	

[a]Dependent variable: TIMSS international math test score. Clustering-robust standard errors in parentheses. Standard errors have schools as the level of clustering unless noted otherwise. Significance levels (based on clustering-robust standard errors): * 1 percent, + 5 percent, ° 10 percent.– [b]Standard error has countries as the level of clustering.

Table A1b: Complete Base Results (Science)[a]

	TIMSS-95		TIMSS-Repeat		Pooled	
	Coef.	*S.E.*	Coef.	*S.E.*	Coef.	*S.E.*
Central exams	0.397*	*(0.099)*[b]	0.359*	*(0.129)*[b]	0.359*	*(0.083)*[b]
Institutional settings						
School responsibility						
School budget	-0.061*	*(0.024)*	-0.012	*(0.022)*	-0.030°	*(0.017)*
Purchasing supplies	0.021	*(0.032)*	0.164*	*(0.033)*	0.101*	*(0.026)*
Hiring teachers	-0.048*	*(0.015)*	0.083*	*(0.018)*	0.011	*(0.012)*
Determining teacher salaries	0.149*	*(0.016)*	0.077*	*(0.017)*	0.136*	*(0.012)*
Teachers' influence						
Strong influence on curriculum						
Teacher individually	0.091*	*(0.015)*	0.082*	*(0.019)*	0.107*	*(0.012)*
Subject teachers	-0.050*	*(0.016)*	0.047+	*(0.019)*	0.001	*(0.013)*
School teachers collectively	-0.045*	*(0.015)*	-0.034°	*(0.018)*	-0.059*	*(0.012)*
Teacher unions	-0.147*	*(0.043)*	-0.353*	*(0.074)*	-0.282*	*(0.055)*
Class teacher has strong influence on						
Money for supplies	0.048+	*(0.021)*	0.043	*(0.031)*	0.039°	*(0.020)*
Kind of supplies	0.025°	*(0.014)*	0.012	*(0.022)*	0.026°	*(0.014)*
Subject matter	-0.021°	*(0.011)*	-0.053*	*(0.016)*	-0.054*	*(0.010)*
Textbook	0.017	*(0.013)*	-0.006	*(0.019)*	0.000	*(0.012)*
Students' incentives						
Scrutiny of testing (hours per week)	0.008+	*(0.004)*	-0.001	*(0.005)*	0.003	*(0.003)*
Homework (hours per week)	-0.013	*(0.008)*	0.000	*(0.010)*	-0.007	*(0.008)*
Parents' influence						
Uninterested parents limit teaching	-0.071*	*(0.025)*	-0.150*	*(0.025)*	-0.142*	*(0.020)*
Interested parents limit teaching	-0.005	*(0.031)*	0.028	*(0.034)*	0.013	*(0.026)*
Centralization						
Central curriculum	0.093	*(0.091)*[b]	0.109	*(0.076)*[b]	0.120	*(0.074)*[b]
Central textbook approval	0.081	*(0.096)*[b]	0.168+	*(0.083)*[b]	0.083	*(0.059)*[b]
Student and family characteristics						
Upper grade	0.454*	*(0.009)*			0.435*	*(0.009)*
Above upper grade	1.056*	*(0.033)*			0.987*	*(0.031)*
Age (years)	-0.083*	*(0.006)*	-0.148*	*(0.007)*	-0.108*	*(0.006)*
Sex (female)	-0.160*	*(0.007)*	-0.156*	*(0.007)*	-0.158*	*(0.005)*
Born in country	0.109*	*(0.013)*	0.268*	*(0.021)*	0.215*	*(0.014)*
Living with both parents	0.048*	*(0.008)*	0.090*	*(0.011)*	0.072*	*(0.007)*
Parent born in country	0.136*	*(0.014)*	-0.032°	*(0.019)*	0.040*	*(0.012)*
Parents' education						
Finished primary	0.193*	*(0.010)*	0.118*	*(0.017)*	0.087*	*(0.009)*
Secondary	0.126*	*(0.011)*	0.299*	*(0.019)*	0.172*	*(0.009)*
Finished university	0.363*	*(0.012)*	0.472*	*(0.020)*	0.357*	*(0.010)*
Books at home						
11-25	0.123*	*(0.012)*	0.158*	*(0.011)*	0.167*	*(0.009)*
26-100	0.334*	*(0.013)*	0.399*	*(0.012)*	0.418*	*(0.010)*
101-200	0.492*	*(0.014)*	0.559*	*(0.013)*	0.574*	*(0.011)*
More than 200	0.588*	*(0.014)*	0.635*	*(0.014)*	0.654*	*(0.011)*
Community location						
Geographically isolated area	-0.088*	*(0.033)*	0.034	*(0.038)*	-0.033	*(0.027)*
Close to the center of a town	-0.015	*(0.012)*	0.084*	*(0.015)*	0.045*	*(0.010)*
GDP per capita (1000 intl. $)	0.040*	*(0.012)*[b]	0.054*	*(0.010)*[b]	0.048*	*(0.008)*[b]

(Table continued on next page)

Table A1b (continued)

	TIMSS-95		TIMSS-Repeat		Pooled	
	Coef.	S.E.	Coef.	S.E.	Coef.	S.E.
Resources and teachers						
Expenditure per student (1000 intl. $)	-0.069	*(0.044)*[b]	-0.123[+]	*(0.046)*[b]	-0.090[*]	*(0.034)*[b]
Class size (no. of students)	0.003[*]	*(0.001)*	-0.006[*]	*(0.001)*	-0.002[*]	*(0.001)*
Student-teacher ratio (10 students)	0.002[*]	*(0.001)*	-0.004[*]	*(0.001)*	-0.002[+]	*(0.001)*
No shortage of materials	0.066[*]	*(0.013)*	0.088[*]	*(0.017)*	0.065[*]	*(0.011)*
Great shortage of materials	-0.092[*]	*(0.020)*	-0.059[*]	*(0.022)*	-0.088[*]	*(0.016)*
Instruction time (100 hours per year)	0.003	*(0.004)*			0.021[*]	*(0.005)*
Instruction time (hours per week)			-0.001	*(0.002)*	-0.005[*]	*(0.002)*
Teacher characteristics						
Teacher's sex (female)	0.089[*]	*(0.011)*	0.084[*]	*(0.015)*	0.101[*]	*(0.010)*
Teacher's age (years)	-0.002	*(0.001)*	-0.001	*(0.001)*	-0.002[+]	*(0.001)*
Teacher's experience (years)	0.004[*]	*(0.001)*	0.007[*]	*(0.001)*	0.007[*]	*(0.001)*
Teacher's education						
Secondary only	0.078[°]	*(0.042)*	-0.271[*]	*(0.083)*	0.069[°]	*(0.038)*
BA or equivalent	0.008	*(0.041)*	0.066	*(0.071)*	0.104[*]	*(0.036)*
MA/PhD	0.154[*]	*(0.043)*	0.177[+]	*(0.072)*	0.247[*]	*(0.037)*
Other post-secondary			0.016	*(0.078)*	0.030	*(0.053)*
Constant	4.355[*]	*(0.107)*	4.475[*]	*(0.170)*	3.977[*]	*(0.102)*
Students (unit of observation)	266545		180544		447089	
Schools (PSUs)	6107		6068		12175	
Countries	39		38		77	
R^2	0.205		0.326		0.256	

[a]Dependent variable: TIMSS international math test score. Clustering-robust standard errors in parentheses. Standard errors have schools as the level of clustering unless noted otherwise. Significance levels (based on clustering-robust standard errors): [*] 1 percent, [+] 5 percent, [°] 10 percent.– [b]Standard error has countries as the level of clustering.

References

Ballou, D. (2001). Pay for Performance in Public and Private Schools. *Economics of Education Review* 20 (1): 51–61.

Barro, R. J. (2001). Human Capital and Growth. *American Economic Review* 91 (2): 12–17.

Baumert, J., W. Bos, and R. Watermann (1999). *TIMSS/III: Schülerleistungen in Mathematik und Naturwissenschaften am Ende der Sekundarstufe II im internationalen Vergleich; Zusammenfassung deskriptiver Ergebnisse*. 2nd revised edition. Berlin: Max-Planck-Institut für Bildungsforschung.

Berliner, D. C. (1990). What's All the Fuss About Instructional Time? In: M. Ben-Peretz and R. Bromme (eds.), *The Nature of Time in Schools: Theoretical Concepts, Practitioner Perceptions*. New York: Teachers College Press.

Bishop, J. H. (1995). The Impact of Curriculum-Based External Examinations on School Priorities and Student Learning. *International Journal of Educational Research* 23 (8): 653–752.

Bishop, J. H. (1997). The Effect of National Standards and Curriculum-Based Exams on Achievement. *American Economic Review, Papers and Proceedings* 87 (2): 260–264.

Bishop, J. H. (1999a). Are National Exit Examinations Important for Educational Efficiency? *Swedish Economic Policy Review* 6 (2): 349–398.

Bishop, J. H. (1999b). Nerd Harassment, Incentives, School Priorities, and Learning. In: S. E. Mayer and P. E. Peterson (1999), *Earning and Learning: How Schools Matter*. Washington, D.C.: Brookings Institution Press.

Bishop, J. H., and L. Wößmann (2001). Institutional Effects in a Simple Model of Educational Production. Kiel Working Paper 1085. Institute for World Economics, Kiel.

Deaton, A. (1997). *The Analysis of Household Surveys: A Microeconometric Approach to Development Policy*. Baltimore, MA: Johns Hopkins University Press.

DuMouchel, W. H., and G. J. Duncan (1983). Using Sample Survey Weights in Multiple Regression Analyses of Stratified Samples. *Journal of the American Statistical Association* 78 (383): 535–543.

Evers, W. M. (2001). Standards and Accountability. In: T. M. Moe (ed.), *A Primer on America's Schools*. Stanford, CA: Hoover Institution Press.

Figlio, D. N., and M. E. Page (2002). Can School Choice and School Accountability Successfully Coexist? Forthcoming in: C. M. Hoxby (ed.), *The Economic Analysis of School Choice*. Chicago, IL: University of Chicago Press.

GEW Hannover (Gewerkschaft Erziehung und Wissenschaft, Hannover) (2002). GEW: Bessere individuelle Förderung statt Selektion und zentraler Prüfungen. Pressemitteilung 31.7.2002, Hannover. www.gew-nds.de/gym/archiv.php [26.9.2002].

GEW Hessen (Gewerkschaft Erziehung und Wissenschaft, Landesverband Hessen) (2002). GEW Hessen lehnt Zentralabitur ab. Presseinfo 19.6.2002, Frankfurt/Main. www.gew-nordhessen.de/Presse/Zentralabi.htm [26.9.2002].

Gonzalez, E. J., and J. A. Miles (eds.) (2001). *User Guide for the TIMSS 1999 International Database*. Chestnut Hill, MA: International Study Center, Boston College, Lynch School of Education, and International Association for the Evaluation of Educational Achievement.

Hanushek, E. A. (1999). Some Findings from an Independent Investigation of the Tennessee STAR Experiment and from Other Investigations of Class Size Effects. *Educational Evaluation and Policy Analysis* 21 (2): 143–163.

Hanushek, E. A. (2002). Publicly Provided Education. Forthcoming in: A. Auerbach and M. Feldstein (eds.), *Handbook of Public Economics, Volume 4*. Amsterdam: Elsevier. (Also available as: NBER Working Paper 8799. National Bureau of Economic Research, Cambridge, MA.)

Hanushek, E. A., and D. D. Kimko (2000). Schooling, Labor-Force Quality, and the Growth of Nations. *American Economic Review* 90 (5): 1184–1208.

Hanushek, E. A. and M. E. Raymond (2001). The Confusing World of Educational Accountability. *National Tax Journal* 54 (2): 365–384.

Hoxby, C. M. (2000). The Effects of Class Size on Student Achievement: New Evidence from Population Variation. *Quarterly Journal of Economics* 115 (4): 1239–1285.

Hoxby, C. M. (2002a). The Cost of Accountability. NBER Working Paper 8855. National Bureau of Economic Research, Cambridge, MA.

Hoxby, C. M. (2002b). Introduction. Forthcoming in: Caroline M. Hoxby (ed.), *The Economic Analysis of School Choice*. Chicago, IL: University of Chicago Press.

Kane, T. J., and D. O. Staiger (2001). Improving School Accountability Measures. NBER Working Paper 8156. National Bureau of Economic Research, Cambridge, MA.

Krueger, A. B. (1999). Experimental Estimates of Education Production Functions. *Quarterly Journal of Economics* 114 (2): 497–532.

Ladd, H. F. (2001). School-Based Educational Accountability Systems: The Promise and the Pitfalls. *National Tax Journal* 54 (2): 385–400.

Ladd, H. F., and R. P. Walsh (2002). Implementing Value-Added Measures of School Effectiveness: Getting the Incentives Right. *Economics of Education Review* 21 (1): 1–17.

Lee, J.-W., and R. J. Barro (2001). Schooling Quality in a Cross-Section of Countries. *Economica* 68 (272): 465–488.

Martin, M. O., K. D. Gregory, and S. E. Stemler (eds.) (2000a). *TIMSS 1999 Technical Report*. Chestnut Hill, MA: International Study Center, Boston College, Lynch School of Education, and International Association for the Evaluation of Educational Achievement.

Martin, M. O., I. V. S. Mullis, E. J. Gonzalez, K. D. Gregory, T. A. Smith, S. J. Chrostowski, R. A. Garden, and K. M. O'Connor (2000b). *TIMSS 1999 International Science Report: Findings from IEA's Repeat of the Third International Mathematics and Science Study at the Eighth Grade*. Chestnut Hill, MA: International Study Center, Boston College, Lynch School of Education, and International Association for the Evaluation of Educational Achievement.

Mauro, P. (1995). Corruption and Growth. *Quarterly Journal of Economics* 110 (3): 681–712.

Moulton, B. R. (1986). Random Group Effects and the Precision of Regression Estimates. *Journal of Econometrics* 32 (3): 385–397.

Mullis, I. V. S., M. O. Martin, E. J. Gonzalez, K. D. Gregory, R. A. Garden, K. M. O'Connor, S. J. Chrostowski, and T. A. Smith (2000). *TIMSS 1999 International Mathematics Report: Findings from IEA's Repeat of the Third International Mathematics and Science Study at the Eighth Grade*. Chestnut Hill, MA: International Study Center, Boston College, Lynch School of Education, and International Association for the Evaluation of Educational Achievement.

Murnane, R. J., and D. K. Cohen (1986). Merit Pay and the Evaluation Problem: Why Most Merit Pay Plans Fail and a Few Survive. *Harvard Educational Review* 56 (1): 1–17.

Organisation for Economic Co-Operation and Development (OECD) (2000). *Education at a Glance: OECD Indicators*. Paris.

Wößmann, L. (2002a). Schooling Resources, Educational Institutions, and Student Performance: The International Evidence. Kiel Working Paper 983, Revised Version, January. Institute for World Economics, Kiel.

Wößmann, L. (2002b). *Schooling and the Quality of Human Capital*. Berlin: Springer.

Wößmann, L. (2003). Specifying Human Capital: A Review and Some Extensions. *Journal of Economic Surveys*: forthcoming.

Kieler Diskussionsbeiträge

Kiel Discussion Papers

382. Familienförderung in Deutschland: Eine Bestandsaufnahme. Von Astrid Rosenschon. Kiel, November 2001. 39 S. 8 Euro.

383. How the EU Can Move to a Higher Growth Path—Some Considerations. By Horst Siebert. Kiel, Dezember 2001. 9 S. 8 Euro.

384. Leviathan in Cyberspace: How to Tax E-Commerce. By Jürgen Stehn. Kiel, Februar 2002. 19 S. 8 Euro.

385. Euroland: Recovery Is Under Way. By Klaus-Jürgen Gern, Christophe Kamps, and Joachim Scheide. Kiel, April 2002. 24 S. 8 Euro.

386. The Stalling Engine in *Wirtschaftswunder-Land:* Germany's Economic Policy Challenges. By Horst Siebert. Kiel, Mai 2002. 16 S. 8 Euro.

387. The European Electricity Market: Centralization of Regulation or Competition between Regulatory Approaches? By Lars Kumkar. Kiel, Mai 2002. 28 S. 8 Euro.

388. IWF und Weltbank: Trotz aller Mängel weiterhin gebraucht? Von Peter Nunnenkamp. Kiel, Mai 2002. 34 S. 8 Euro.

389./ 390. Fit für die EU? Indikatoren zum Stand der Wirtschaftsreformen in den Kandidatenländern. Von Federico Foders, Daniel Piazolo und Rainer Schweickert. Kiel, Juni 2002. 69 S.16 Euro.

391. Fortschritte beim Aufbau Ost. Forschungsbericht wirtschaftswissenschaftlicher Forschungsinstitute über die wirtschaftliche Entwicklung in Ostdeutschland. Kiel, Juni 2002. 53 S. 8 Euro.

392./ 393. Subventionen in Deutschland. Von Alfred Boss und Astrid Rosenschon. Kiel, August 2002. 71 S. 16 Euro.

394. 75 Punkte gegen die Arbeitslosigkeit. Von Horst Siebert. Kiel, August 2002. 23 S. 8 Euro.

395. Vom Mangel zum Überfluss – der ostdeutsche Wohnungsmarkt in der Subventionsfalle. Von Dirk Dohse, Christiane Krieger-Boden, Birgit Sander und Rüdiger Soltwedel. Kiel, September 2002. 52 S. 8 Euro.

396. Euroland: Upswing Postponed. By Kai Carstensen, Klaus-Jürgen Gern, Christophe Kamps and Joachim Scheide. Kiel, Oktober 2002. 19 S. 8 Euro.

397. Central Exams Improve Educational Performance: International Evidence. By Ludger Wößmann. Kiel, Oktober 2002. 45 S. 8 Euro.

Mehr Informationen über Publikationen des Instituts für Weltwirtschaft unter http://www.uni-kiel.de/ifw/pub/pub.htm, mehr Informationen über das IfW unter http://www.uni-kiel.de/ifw/

Institut für Weltwirtschaft an der Universität Kiel, 24100 Kiel
Kiel Institute for World Economics