Working Paper No. 199

Central Exams and Adult Skills: Evidence from PIAAC

Lisa Leschnig, Guido Schwerdt and Katarina Zigova

Universität Zürich
IBW – Institut für Betriebswirtschaftslehre

UNIVERSITÄT BERN

Working Paper No. 199

Central Exams and Adult Skills: Evidence from PIAAC

Lisa Leschnig, Guido Schwerdt and Katarina Zigova

July 2022

Published as: “Central Exams and Adult Skills: Evidence from PIAAC.“ *Economics of Education Review*, 90(2022). By Lisa Leschnig, Guido Schwerdt and Katarina Zigova.

DOI: https://doi.org/10.1016/j.econedurev.2022.102289

Central Exams and Adult Skills: Evidence from PIAAC

Lisa Leschnig* Guido Schwerdt†§ Katarina Zigova‡

July 14, 2022

Abstract

Central exams are often hypothesized to favorably affect incentive structures in schools. Indeed, previous research provides vast evidence on the positive effects of central exams on student test scores. But critics warn that these effects may arise through the strategic behavior of students and teachers, which may not affect human capital accumulation in the long run. Exploiting variation in examination types across school systems and over time, we provide the first evidence that central exams positively affect adult skills. However, estimated effects on skills are small and we find no significant average effect on earnings.

JEL-Code: I20, J24, J31
Keywords: Central exams, adult skills, earnings, PIAAC

We thank Uschi Backes-Gellner, Rick Hanushek, Jeff Smith, Simon Wiederhold, Ludger Woessmann, Stefan Wolter, and seminar participants at the universities of Konstanz, Maastricht, Zurich, and Bochum, the Australian National University (ANU), the KU Leuven, the CESifo meeting on the Economics of Education, the European Society of Population Economics, the International Conference on PIAAC and PIAAC-Longitudinal in Mannheim, the Spring Meeting of Young Economists, as well as participants at the annual meeting of Education Economists in the German Economic Association for valuable discussion and comments. We also thank two anonymous reviewers for helpful comments on an earlier draft of the manuscript.

*Federal Ministry of Education and Research Germany (BMBF); The views expressed in this article are those of the authors and do not necessarily represent the position of the BMBF.

†University of Konstanz, IZA, CESifo, ROA

‡University of Kassel

§Corresponding author. E-mail address: guido.schwerdt@uni-konstanz.de

1 Introduction

Central exams are associated with substantially higher test outcomes of students (see Bishop, 1997; Woessmann, 2003, 2005; Jürges et al., 2005; Fuchs and Woessmann, 2007). While this reduced-form pattern is well documented, critics warn that this result may simply reflect differences in students' test-taking ability, rather than actual differences in knowledge and skills. If that were so, central exams might not genuinely improve the productive skills of adults. Indeed, the existing evidence on the relationship between central exams and labor market outcomes is at best mixed (Bishop et al., 2000; Backes-Gellner and Veen, 2008; Piopiunik et al., 2013). However, there is no evidence on the direct effect on cognitive skills of adults, and the available evidence on the long-term impacts of central exams on labor market outcomes rests entirely on observational studies that estimate cross-sectional regression models.

This is the first paper to study the relationship between the type of exit examination at the end of secondary school and the cognitive skills of adults, as well as labor market outcomes in a panel framework. For this purpose, we draw upon data from OECD's survey of adult skills (OECD, 2013b). Internationally, the Programme for the International Assessment of Adult Competencies (PIAAC) is the most comprehensive survey of adult skills ever undertaken. It assesses literacy and numeracy skills and the ability to solve problems in technology-rich environments. It also collects information on labor market outcomes and background characteristics of representative samples of respondents aged 16-65. We supplement the PIAAC data with data on reforms of examination regimes since 1960 for 30 countries participating in PIAAC. We collected these data specifically for this analysis from a wide variety of sources (cf. Table A1). The data allow us to assign to each participant of PIAAC the type of exit exam that was in place at the time when the participant graduated from high school. Our classification of exam regimes focuses on the central versus local nature of high school leaving exams and employs the definition of central exams of Bishop (1997).

We leverage the panel structure (by graduation cohort across countries) of our data to identify the long-term effects of central exams by exploiting within-country variation in exam types over time and across countries.[1] In particular, we estimate regression models that include country fixed effects as well as country-specific linear trends, which ensure that unobserved time-invariant differences in relevant characteristics or differ-

[1]Note that exploiting this variation is different from comparing U.S. states with and without minimum competency exams or differing graduation requirements (e.g., Baker and Lang, 2013; Bishop and Mane, 2001; Dee and Jacob, 2007), as here graduation requirements in many cases do not change.

ences in linear trends across countries do not confound our estimates. To benchmark our panel estimates, we also present results based on commonly estimated cross-sectional models that rely entirely on strong selection-on-observable assumptions.

Our findings reveal a positive effect of central exams at the end of secondary school on adult skills. We find that central exams on average increase cognitive skills of adults by about 7 percent of a standard deviation. Effect sizes are similar for numeracy and literacy, and only slightly higher for problem-solving skills. We find no significant effects on earnings, employment, or the probability of holding a tertiary degree. In terms of earnings, we can rule out effect sizes larger than 4.6 percent.

Our estimated average effects on skills are small compared to the results of studies that examine student skills, which typically find central exam effects of more than 10 percent of a standard deviation. This difference might be the consequence of some fade-out in central exam effects. Given recent estimates of the returns to adult skills based on the same data (Hanushek et al., 2017), our small estimate of the skill effect is consistent with not finding any earnings effects above 4.6 percent.

However, we also find evidence for effect heterogeneity across countries and population groups. In particular, while all groups benefit from central exams, we find some evidence for stronger effects on skills for second-generation migrants and also some evidence for positive earnings effects for females and individuals with low parental education background. Most interestingly, we also find some evidence suggesting that central exam effects positively depend on the level of school autonomy in a country. Overall, these findings are in line with previous evidence on heterogeneities of central exams on student achievement (see Woessmann, 2005).

We also conduct a series of robustness checks that reveal some interesting aspects. First, our main results are robust to a large set of sensitivity checks. In particular, we present a decomposition of our twoway fixed effects estimate as recently proposed in Goodman-Bacon (2021) and additionally report results based on Callaway and Sant'Anna (2021)'s doubly robust difference-in-differences estimator to show that our twoway fixed effects estimates of the average reform effect are not biased because of time-varying treatment effects. Second, we supplement our main analysis with comparable estimates of the central exam effect on student skills. While these results are purely descriptive, they support the conjecture that central exam effects on skills fade out to some extent over time. Finally, we also provide some evidence for the existence of a positive relation between central exams and adult skills by exploiting the variation in exam types across German federal states. This finding within a single country concurs with our international results.

The remainder of the paper is structured as follows. Section 2 reviews existing evidence on the effects of central exams. In Section 3, we describe our self-collected longitudinal data on exit examinations, the PIAAC data as well as the construction of our estimation sample. In Section 4, we present our empirical strategy. Section 5 presents our main results, Section 6 discusses their robustness, and Section 7 presents further supporting evidence. Our conclusion is in Section 8.

2 Conceptual background

We begin our discussion with a definition of central exams, based on Bishop (1997).[2] A central school leaving examination (in short, a central exam) is a mandatory written test, administered by a central authority (such as a ministry of education), that provides centrally developed, curriculum-based test questions covering core subjects such as math and the national language. Central exams impose consistent test implementation, grading, and a pre-defined passing threshold. Thus, central exams set external and consistent quality standards of secondary education (not necessarily skills), and ensure comparability of students' performance. Additionally, achievement in a central exam—and, especially, failing a central exam—is required to have direct consequences for students who take the exam, for example by representing a significant part of the final grade and by a limited number of valid retakes. Finally, central exams can be organized either on a national or regional level.[3]

The theoretical foundation for our empirical analysis of the effects of central exams is laid out in models such as Bishop and Woessmann (2004) and Bishop (2006). Central exams may affect student skills and later outcomes through a variety of channels. First, by providing information on the outcomes of the educational process, central exams can improve the monitoring of the behavior of teachers and schools. Thus, central exams may raise educational outcomes as they improve accountability in school systems. Second, if central exams improve the signal of educational achievement for employers and institutions of higher education (Becker and Rosen, 1992; Schwerdt and Woessmann, 2017), they can have true productivity effects by creating incentives for students to

[2]See Bishop (1997, pg. 260) for the detailed definition of a central exam.

[3]Note that, according to the definition of Bishop (1997), commercially prepared tests such as the Scholastic Assessment Test (SAT), American College Testing (ACT), or the California Achievement Test (CAT) do not classify as central exams. These tests fail to have direct consequences for students since students can retake them multiple times and the test content is neither curriculum-specific, nor are these tests mandatory for graduation.

exert more effort.[4] Hvidman and Sievertsen (2021) provide evidence that grading incentives in high stakes tests significantly impact students' learning effort. Exploiting a school reform that endogenously re-coded the GPA of high school students, the authors show that students whose prior achievement was downgraded exerted a greater effort in the following school years to compensate for the reform-induced demotion. Finally, central exams may also decrease collective peer pressure against learning, because they render futile any collective strategy to lower standards in a classroom—which again increases learning outcomes.

An extensive empirical literature has investigated the effects of central exams on students' skills. Evidence from international student achievement tests indeed shows that students perform substantially better in countries with central exams than in countries without them (see Hanushek and Woessmann, 2011, for an overview). Studies like Bishop(1997, 1999a), Woessmann (2003), Woessmann (2005), Bishop (2006) or Fuchs and Woessmann (2007) suggest that the effects of central exams on student achievement may well be larger than a whole grade-level equivalent, or between 20 and 40 percent of a standard deviation of the respective international tests. Studies in countries with regional variations in exam systems find similar results.[5]

To probe causality, Jürges et al. (2005) apply a differences-in-differences approach to the German TIMSS 1995 data that exploits the fact that in some secondary-school tracks, the states with central exams have exams in math but not in science, finding a smaller, but still substantial, effect of about 0.13 of a national standard deviation. More recently, Bergbauer et al. (2021) show that an expansion in standardized external comparisons is associated with improvements in student achievement. Their panel analysis is based on data on students observed over six waves in the international PISA student achievement test 2000–2015, and exploits reforms in assessment systems over time for identification. Most importantly, they show that both school-based and student-based external comparisons have positive effects on student achievement, where their definition of a student-based external comparison is closely aligned with our definition of a central exam.[6]

But students' higher test scores, as induced by central exams, may not automatically translate into better outcomes for adults. More school accountability and binding

[4] In a similar spirit, principal-agent models of educational standards model how educational credentials affect the level of learning effort that students choose (e.g., Costrell, 1997; Betts, 1998).

[5] Positive effects of central exams on test-score outcomes have been shown for Canadian provinces (Bishop, 1997, 1999a), for U.S. states (e.g., Bishop et al., 2000), and for German states (Jürges et al., 2005; Jürges and Schneider, 2010; Jürges et al., 2012; Lüdemann, 2011).

[6] In fact, Bergbauer et al. (2021) use our compiled data on central exams at the end of secondary school to create their measure of student-based external comparisons.

standards through external testing could also have negative effects (see Figlio and Loeb, 2011, for a review). Critics often argue that test-based accountability systems may raise only test-taking skills, not genuine educational achievement (Popham, 2001; Koretz, 2002; Volante, 2004). Increased test scores can also just be due to fraudulent behavior such as outright cheating of teachers(Jacob and Levitt, 2003). Moreover, studies for the United States (e.g. Figlio and Getzler, 2006; Cullen and Reback, 2006) show that under educational accountability policies, schools tend to strategically place low-performing students in special education tracks or bilingual programs that are not subject to central testing to increase school performance. In addition, Jürges and Schneider (2010) find that while central exams do improve students' academic skills, they negatively affect students' attitudes toward learning, as indicated in self-reported enjoyment of mathematics. And Jürges et al. (2012) indicate that the positive effect of central exams on curriculum-based knowledge does not extend to a positive effect on mathematical literacy.[7]

Central exams may also cause heterogeneous performance effects, depending on students' backgrounds or other characteristics of education systems. Central exams could discourage especially low performing students who see little chance to pass a central exam (Bishop, 1999b) and higher accountability standards could amplify the achievement gap for minority students because of perceived negative stereotypes (Dee and Jacob, 2007). In a cross-sectional analysis of the effect heterogeneity of central exams, Woessmann (2005) finds that central exam effects overall vary only little along most dimensions of student background, but also provides some evidence that students with an immigrant or less-educated family background seem to benefit relatively more from central exams. Interestingly, Woessmann (2005) finds evidence for a substantial heterogeneity in the effect of central exams depending on whether schools have autonomy. The economic rationale for this complementarity between central exams and school autonomy is explained in the context of a simple principal-agent model. In such a model the decentralization of decision-making power can be understood as the delegation of a task by a principal (e.g., an education ministry) to agents (schools) in order to capitalize on the better local knowledge of agents. In this model, central exams can reduce the danger of opportunistic behavior of autonomous agents by providing a monitoring instrument.

Ultimately, any changes in students' test scores, as induced by central exams, may not automatically translate into changes in outcomes for adults. Only a few studies

[7] Jürges et al. (2012) distinguish between test achievements reflecting the German mathematics curriculum and achievements in applying the knowledge to real world issues (i.e., mathematical literacy).

have investigated this link. Backes-Gellner and Veen (2008) and Piopiunik et al. (2013) study labor market outcomes exploiting cross-sectional variations in exit exams across German federal states. Backes-Gellner and Veen (2008) fail to find a positive effect of central exams on labor-market earnings for graduates of the highest school track of the highly tracked German school system. But Piopiunik et al. (2013) show that graduates from the lowest track have higher earnings if they received their high school leaving certificate in a state with central exams. Piopiunik et al. (2013) also show that graduates from both low-track and high-track schools have lower rates of unemployment when their school exit exam was centrally administered. However, these findings may be specific to Germany's rigid labor market, where earnings structures are mostly determined by central bargaining.

For the United States, Bishop et al. (2000) provide evidence on the effects of central exams, but their analysis ultimately boils down to a comparison of New York State to the remaining states. Wider variation across U.S. states is restricted to course graduation requirements and minimum competency exams, which assess only low-level skills in public schools and have no consequences for university entrance. Based on longitudinal data that allows linking the exam type of individual students with later labor-market outcomes, Bishop and Mane (2001) find minimum competency exams, but not mere course graduation requirements, to be positively associated with earnings.

Our paper directly contributes to this literature on central exams by providing a first analysis of the direct effect of central exams on adult skills, and a first analysis of the effect on labor market outcomes, in a panel framework that exploits several reforms to examination systems across a wide range of countries. We believe that this extension is important for several reasons. First, the existing evidence on the effect on labor market outcomes is based on regression models that exploit variation in exam regimes across regions. The evidence is therefore prone to biases arising because of unobserved institutional or cultural differences across regions. Second, the existing evidence on labor market effects is also limited to specific countries and might not be generalizable. Third, studying the direct effect of central exams on adult skills is important because positive effects of centralized examination systems may also arise through other channels besides increased human capital accumulation. For example, central exams may facilitate the matching process between vacancies and workers, as it allows for an improved sorting of students by productive skills. This may increase overall productivity in the long run.

3 Data

In this section we describe our data collection and the construction of our estimation sample. We also elaborate on how we match our data on exit examinations in 30 countries to individual-level data of the Programme for the International Assessment of Adult Competencies (PIAAC).

3.1 Longitudinal data on central exams

For this analysis, we make use of a large variety of sources to collect data on reforms of exit exams at the end of secondary school in all 30 countries for our study period from 1960 to 2012 (2015 for PIAAC round 2 countries).[8] Our sources include international and national publications, both online and offline sources, and personal communications with country experts. We provide a detailed documentation of our country-specific sources in Appendix—Table A1. Based on this collected data on the history of exam regimes in all 30 countries, we classify for each graduation cohort whether it was exposed to a high school leaving exam that qualifies as a central exam according to the definition provided in Bishop (1997).

Table 1 shows the central exam status and, if applicable, reform years between 1960 and 2015 across the 27 countries in our sample with no variation in exam regimes at the sub-national level. Reform years indicate the first cohort of high school graduates affected by the reform. In 9 countries central exams were introduced before 1960, while in 5 countries no central exams were in place during our study period. In 13 countries we identified relevant changes in the central exam status at the national level during our study period. In most cases, these changes were education reforms that introduced central exam. However, we also observe the abolishment of central exams at the national level. Sweden had central exams until 1968 and Greece first introduced central exams in 1983 but abolished them again in 2005. We therefore observe a total of 14 changes in the central exam status at the national level in our data.

In addition to these reform countries with national education systems, we also have 3 reform countries with federal education systems in our sample in which we observe changes in the central exam status at the level of provinces or federal states within the countries. These countries are Canada, Germany, and the United States. Table A2

[8] This analysis includes countries from the first two rounds of data collection in the first cycle of PIAAC. In late 2019, data on 5 additional countries became available in a third round of data collection. Further analysis based on a preliminary classification of the history of exam regimes at the end of secondary education in these 5 additional countries suggests that all results are robust to the inclusion of these countries in the estimation sample. Results including these 5 additional countries are available from the authors upon request.

provides information on exit exams and on reforms affecting the central exam status in these countries over time. In total, we observe 25 changes in the central exam status at either the province or the state level within these 3 countries in our study period.

There are three peculiarities about our coding of exam regimes that are important to note. First, while the treatment status of having a central exam is well defined, the counterfactual is not. School systems may not satisfy our definition of having a central exam for a variety of reasons. On the one hand, there might be a general absence of any school leaving exam (as in Spain and Belgium). On the other hand, some school leaving exams violate our definition across one or more dimensions. For example, exit exams could be administered at the school district or school level (as in Greece and some federal states in Germany), may not be mandated for graduation (Sweden after 1968), may fail to have direct consequences for graduates (perhaps because of the option of multiple retakes, as with CAT and SAT tests or in most of the U.S. graduation exams), or may not be curriculum-based (as in many U.S. states).

Second, we generally distinguish between university entrance exams and high school leaving exams, as only students who are aiming to attend a university typically participate in university entrance exams. However, previous studies on central exams (e.g., Bishop, 1997; Fuchs and Woessmann, 2007) treated university entrance exams in Japan and Korea as central exams, as in both countries the overwhelming majority of high school graduates takes part in the centrally organized university entrance exams. To stay consistent with the literature, we also adopt this view in our preferred specification and treat university entrance exams in Korea and Japan as central exams. However, in a specification where we exclude these two countries from our analysis we also show that our main results stay robust.

Third, we ultimately cannot fully exploit the sub-national variation in central exams across regions over time in the 3 reform countries with federal education systems as the PIAAC data does not provide information on the province or state of residence for surveyed individuals. Instead we follow the common practice in empirical studies on central exams and assign a central exam factor to graduation cohorts in countries with education systems organized at a sub-national level (for example, Fuchs and Woessmann, 2007).[9] For this purpose, we weight regional central exam dummies for each state and each graduation year with the state's respective population share in 2012. Aggregating the weighted dummies, we obtain a single factor at the country level. For a given graduation cohort and on the federal level, this factor should roughly reflect

[9]Germany is an exception, as we can identify the federal state of residence in the German data. We exploit this information in Section 6.4.

the probability that an individual was exposed to a central exam regime.

Figure 1 displays the total variation in the central exam status over time in our full sample of 30 countries. This variation includes the changes in the dichotomous classification of central exams in the 13 reform countries with national education systems as well as the evolution of the factors for the three reform countries with federal education systems. These factors range in Canada from 0.32 to 0.57, in Germany from 0.35 to 0.95, and in the United States from 0 to 0.063.

3.2 PIAAC data

The Programme for the International Assessment of Adult Competencies, was developed by the OECD to provide internationally comparable data on adult skills (OECD, 2013a; OECD, 2016). The first round of PIAAC was administered to 22 countries between August 2011 and March 2012; the second was administered to additional 8 countries between April 2014 and March 2015. The two rounds combined thus provide comparable skill data for 30 countries.[10] In each participating country, a representative sample of at least 5,000 adults between 16 and 65 years of age was interviewed at home in the language of their country of residence. The standard survey mode was to answer questions on a computer, but respondents without sufficient computer knowledge could also do a pencil-and-paper survey.

PIAAC was designed to measure key cognitive and workplace skills needed for individuals to advance at work and participate in society. The survey assessed cognitive skills in three domains: numeracy, literacy, and problem-solving in technology-rich environments.[11] The tasks that respondents had to solve were often framed as real-world problems, such as maintaining a driver's logbook (numeracy domain) or reserving a meeting room on a particular date using a reservation system (problem-solving domain). The domains are described in more detail in OECD (2013a). For analytical purposes, we standardize skills in the full sample to have mean zero and standard deviation (SD) one, and employ the sample weights provided in PIAAC. In the empirical analysis, we focus on numeracy skills, which we deem most comparable across countries, but we report the results for the other two skill domains as well.

[10]The actual number of countries surveyed in the first and the second PIAAC rounds was 24 and 9. Following the OECD recommendation, we excluded Russia because the Russian sample does not cover the population of the Moscow region. We also had to exclude Australia because of restrictive access to the Australian data. From the second-round countries, we excluded Indonesia as it withdrew its participation in the PIAAC study and submitted only data from the Jakarta region.

[11]Participation in the problem-solving domain was optional; Cyprus, France, Italy, and Spain did not participate in this domain.

We also investigate whether central exams affect labor market and education outcomes. For this analysis, we make use of the data on labor-market status, earnings, education, and demographic characteristics, as provided by PIAAC. Specifically, we estimate central exam effect on gross hourly earnings of wage and salary workers, employment status, and educational attainment. These outcome measures have been used previously in the literature analyzing PIAAC data (for example Hanushek et al., 2015).[12]

We match our self-collected data on exam regimes to the PIAAC data by country of residence and year of completion of the highest level of schooling. That is, we assign to each PIAAC participant the type of exit examination at the end of high school that was in place in her country at the time of her graduation. High school graduation in the PIAAC data is defined as having completed at least the Level 3a of the International Standard of Classification of Education (ISCED), according to the 1997 classification. As PIAAC provides information on the highest educational degree obtained and the age at which individuals left full-time education, we can derive the year of graduation directly for individuals whose highest educational degree is the high school diploma. Individuals who report having not graduated from high school are kept in our baseline sample and are assigned to the graduation cohort corresponding to the year they report having left full-time education. For individuals with higher educational degrees, we assume their graduation age to be the country-specific median age at high school graduation among those with high school as their highest degree. To ensure that all individuals in our sample have been exposed to at least two years of secondary schooling, we restrict our baseline sample to individuals with a graduation age of 15 or above.[13]

Note that our matching procedure implies a specific treatment definition, as some individuals in our sample will be considered to have been exposed to a central exam regime even though central exams were only introduced in their final high school year. Thus, some students whom we consider to be treated, might not themselves have participated in a central exam. However, they may be indirectly affected by the change in exam regimes (such as by spillover effects of increased school accountability). As a consequence, the treatment we are studying here could be described as "*at least some (direct or indirect) exposure to a central exam regime.*"

[12]In particular, see Hanushek et al. (2015) for detailed information on the construction of our earnings measure.

[13]Results are, however, almost identical when early school leavers are not excluded from the sample. See Table 5, "Incl. early leavers" row .

Table 2 provides summary statistics for our baseline sample in all 30 countries by exam regime. The table shows that, on average, individuals exposed to central exams are quite different from those in non-central exam regimes. In particular, they are more skilled, earn more, are more likely to be employed, and are more likely to hold a tertiary degree. These differences are not surprising, as the younger population is over-represented in the central exam category.

4 Empirical strategy

We estimate the effect of central exams on adult outcomes based on regression models of the following kind:

$$y_{icg} = \beta\, CE_{cg} + X'_{icg}\gamma + \delta_c + \theta_g + \lambda_c g + \varepsilon_{icg}, \tag{1}$$

where y_{icg} is the outcome of individual i in country c belonging to graduation cohort g, CE_{cg} is a dummy that indicates whether an individual belonging to graduation cohort g in country c was exposed to a central exam, δ_c is a country fixed effect, θ_g is a graduation cohort fixed effect, $\lambda_c g$ is a country-specific linear trend in graduation years, X_{icg} is a vector of student background variables, and ε_{icg} is an idiosyncratic error. When presenting our results, we report estimates of this model with and without country-specific linear trends in graduation years. To benchmark our panel estimates, we also estimate cross-sectional versions of this model that do not control for any unobserved country-specific effects.

The key parameter of interest in Equation 1 is β. It captures the effect of graduating in a central as opposed to non-central exam regime. By leveraging the panel nature of our data at the cohort-country level, we identify β in our preferred model solely based on within-country variation in examination regimes across graduation cohorts.[14] In particular, our identification strategy makes use of the fact that some countries changed their exam regimes in different years between 1960 and 2012 (2015 for PIAAC round 2 countries), while others did not. Thus, unobserved institutional or cultural differences across countries that have a constant impact on adult outcomes do not bias our estimates.

In our full sample, our estimation approach identifies an average effect over a total of 39 education reforms (14 reforms at the national and 25 reforms at the sub-national level) that changed the central exam status of a country. The model in Equation 1

[14]Other studies that exploit variation in the exposure to education policies across cohorts in cross-sectional data for identification include Pischke (2007) and Pischke and von Wachter (2008).

is a twoway fixed effects model with country-specific linear trends and can be seen as difference-in-differences (DD) design with differential timing of treatment. Recent work has shown that the twoway fixed effects estimate, $\widehat{\beta}$, is a weighted average of all potential 2×2 DD estimates where weights depend on group sizes and variance in treatment (see Goodman-Bacon, 2021). It identifies a meaningful average reform effect on the treated under the assumption of time-invariant treatment effects and variance weighted common trends.[15] While this assumption can ultimately not be tested, graphical evidence from an event study analysis presented in Figure A1 reassuringly reveals no systematic difference in pre-trends between reform countries and countries that never adopted central exams.[16]

We conduct a wide range of sensitivity analysis to check on the causal interpretation of our estimates. In particular, we check whether our results are sensitive to allowing countries to follow different linear trends, which is commonly seen as an important specification check in the context of difference-in-differences designs (see Angrist and Pischke, 2015, p.199).[17] In Section 6 we conduct further sensitivity checks including a permutation test to assess whether our estimates are likely to be driven by accidental correlations between changes in exam regimes and other unobserved discrete changes that impact skill development. Moreover, we show results of a decomposition exercise proposed by Goodman-Bacon (2021) to assess the potential of time-variant treatment effects affecting our estimate and report estimates of the average treatment effect on adult skills based on the doubly robust estimator developed in Callaway and Sant'Anna (2021). In Section 7 we probe the heterogeneity of effects using different subgroups of data defined by population characteristics and by the level of autonomy of schools in different education systems.

Throughout the empirical analysis, we estimate and report standard errors clustered at the country by graduation cohort level. Our primary motivation for clustering standard errors at this level is design-based, as our treatment (the exposure to a central exam) is constant within these clusters. We also follow other cross-country studies based on PIAAC data (such as Hanushek et al., 2015) and estimate weighted regres-

[15] Goodman-Bacon (2021) shows that the assumption of variance weighted common trends is sufficient in this setting and that it is slightly weaker than the common assumption of exact parallel trends.

[16] Details on the event study analysis are reported in the note to Figure A1. Note, that all leads are jointly insignificant for numeracy and problem solving, but are marginally significant for literacy. All lags are significantly different from zero for all three skill domains. However, Figure A1 suggest that treatment effects might be dynamic. A point we explore in more detail in Section 7.2.

[17] For example, Stephens and Yang (2014) have shown, in the context of studies exploiting reforms to compulsory schooling laws for identification, that identification based on the timing of policy changes across states over time can be very sensitive to the inclusion of state-specific trends.

sions using sampling weights, which we adjusted so that each country is given the same weight to account for different sample sizes across PIAAC countries.

5 Results

All else being equal, do central exams at the end of secondary education affect adult outcomes? In addressing this question, we begin with a set of estimates of the effects on adult skills, before presenting estimates of the effects on labor market and educational outcomes.

5.1 The effect of central exams on adult skills

We report our main estimates of the effects of central exams on adult skills in Table 3. We start by presenting a set of benchmark estimates that do not control for unobserved heterogeneity across countries in columns 1 and 2 and follow by presenting estimates of our preferred model given by Equation 1 in columns 3 to 8. As skill measures are standardized to have a mean of zero and standard deviation of one, the coefficient of interest can be interpreted as the effect of central exams on adult skills measured in percent of a standard deviation of the respective skill distribution.

Across the different specifications, there is clear evidence that central exams are positively related to adult skills. Without taking into account any control variables and unobserved heterogeneity across countries, column 1, graduates in central exam regimes outperform graduates in non-central exam regimes by about 20 percent of a standard deviation in the numeracy test. The estimate decreases slightly to about 17 percent in column 2 when controlling for individual background characteristics and the age at graduation. These effect sizes are comparable to estimates of central exam effects on student skills based on cross-sectional data (for example Bishop, 1997; Woessmann, 2002; Fuchs and Woessmann, 2007). The estimates on the background variables reveal several typical features of estimated education production functions that are in line with previous evidence.[18]

Turning to our panel estimates, we observe that estimates become much smaller in absolute terms, but remain positive and significant. Including country fixed effects in column 3 suggests that central exams increase numeracy skills on average by about 9 percent of a standard deviation. Column 4 introduced country-specific linear

[18] In line with estimates of education production for students (e.g., Fuchs and Woessmann, 2007), we also find that skills are positively related with parental education background, negatively related with migration background, and a more pronounced gender gap in favor of males in numeracy than literacy.

time trends, which allows us to control, to some extent, for unobserved country-specific confounders that have a different impact on outcomes across graduation cohorts. Introducing these country-specific trends decreases the estimate only slightly to 7 percent. For literacy skills, estimated effects in columns 5 and 6 are similar to those for numeracy, while the estimated effect on problem-solving skills, in column 8, is somewhat larger, at 9 percent of a standard deviation.

Our estimates on adult skills are about half the size of well-identified estimates of central exam effects on student skills (such as Jürges et al., 2005). This may indicate that the effect of central exams fades out as individuals grow older, which may reflect that part of the initial effect on student test scores is indeed driven by an increased test-taking ability gained during secondary education by students in central exam regimes.

5.2 The effect of central exams on labor market outcomes

To investigate whether the increase in adult skills induced by central exams also translates into better labor market outcomes, in Table 4 we report estimates of the central exam effects on earnings, employment, and college degree. The first column in Table 4 reports coefficient estimates of variants of Equation 1, with individual earnings as dependent variable. In column 2 we report estimates of the effect on employment.[19] Finally, we test whether central exams affect the probability of completing tertiary education and receiving a high school diploma in columns 3 and 4, respectively.[20]

Once we control for unobserved country-specific trends, we do not find any significant effects on these outcomes. Most importantly, our estimates let us rule out the existence of earnings estimates larger than 4.6 percent.[21] This finding is, however, consistent with the finding of effects on adult skills of about 7 percent of a standard deviation if central exams affect earnings primarily through a skill production channel. Hanushek et al. (2017) report an average wage return to a one-standard-deviation increase in numeracy skills of 20 percent. Based on a back-of-the-envelope calculation, we should in this case expect that a shift in skills of 7 percent of a standard deviation would produce a shift in earnings of only about 1 percent. Interestingly, our point estimate actually suggests an effect on earnings of about 1 percent, but we lack the statistical power to detect such small effect sizes.[22]

[19] Any participant in PIAAC who reports positive labor earnings is coded as employed.

[20] Note that in columns 3 and 4, we follow Hanushek et al. (2015) and exclude workers below age 25, as they are not likely to have completed their first phase of full-time education.

[21] The upper bound of the 95 percent confidence interval of the earnings estimate in the first column of Table 4 is 0.0461.

[22] The calculation is as follows: Central exams positively affect skills by about 7 percent of a standard

6 Robustness

In this section we conduct a set of sensitivity checks to assess the robustness of our main results.

6.1 Sensitivity checks

We begin by reporting estimates of a series of sensitivity checks in Table 5. To consolidate our presentation, we report estimates of the central exam effect based on six outcomes in the first row. As shown in the second row of Table 5, our results prove to be mainly robust to estimating regression models that do not use any survey weights. Another robustness concern may be that the graduation year is partly endogenous to the type of exit exam. We test the relevance of this concern by making an alternative assumption about the graduation age of individuals. Estimations in the third row are based on assigning the median graduation age within a country to all graduates in the country. The results show the robustness of our main results to this alternative assumption about individuals' graduation age. Finally, we show that our main results are generally robust to the inclusion of individuals who report having left school before age 15.

Another key concern is that our overall finding is entirely driven by a specific country or a small set of countries. To investigate this possibility, we estimate our preferred specification several times, each time excluding a different country or group of countries. Table 6 presents the results from this piecewise deletion exercise. The results are very similar throughout, which shows that our main results are not driven by any specific country or a small set of countries. Note, however, that we observe some variation in the estimated average effect of central exams on skills if we selectively drop reform countries with small (Canada, Greece, Japan, and U.S.) and large (Germany, Poland, and Singapore) country-specific reform effects. This indicates some degree of heterogeneity of the estimated reform effects across countries.

Finally, we also report standard errors clustered at the country level in square brackets for all baseline estimates in the first row of Table 5. Clustering standard errors at the level of countries can improve inference in case of intertemporal correlation of errors within countries. With too few clusters, however, clustered standard errors might also not be very reliable.[23] Fortunately, our result are largely robust to clustering at the

deviation, while a standard deviation increase in skills increases earnings by about 20 percent (see Hanushek et al., 2017)). Thus, the expected effect of central exams on earnings is 0.07×20, which comes to 1.4 percent.

[23] In recent work Abadie et al. (2022) also argue that clustered standard errors at more aggregate levels may be unnecessarily conservative in settings with variation in the distribution of the treatment across clusters.

country level. Effects on labor market and education outcomes are still insignificant and the somewhat larger effect on problem-solving skills remains significant in our baseline specification with country-level clustering. Only the estimated coefficients on numeracy and literacy in our baseline specification with country-specific linear trends become insignificant (p-values of 0.14 and 0.22) with country-level clusters.[24]

6.2 Staggered introduction of reforms

A recent literature documents possible shortcomings of two-way fixed effects specifications with time and state fixed effects in the form of equation (1) (Callaway and Sant'Anna, 2021; De Chaisemartin and d'Haultfoeuille, 2020; Goodman-Bacon, 2021; Sun and Abraham, 2021).

In particular, Goodman-Bacon (2021) shows that a twoway fixed effects estimate is a weighted average of all potential 2×2 DD estimates. These different groups of canonical 2×2 DD estimates also include estimates in which earlier-treated countries are effectively serving as controls for later-treated countries. Such comparisons are, however, problematic in the presence of time-varying treatment effects.

To show the anatomy of our estimate of average central exam effects, we implement the decomposition exercise suggested by (Goodman-Bacon, 2021) based on a reduced sample of countries that allows a straightforward application of the method.[25] Table 7 summarizes the findings of the decomposition exercise, while Appendix Figure A2 visualizes the variation in the 2×2 DD estimates from the Goodman-Bacon decomposition and their weight for analysis of the average central exam effects.[26] Note, that estimates of the effects on adult skills are all close to the corresponding estimates based on our full sample of countries reported in columns 3, 5, and 7 of Table 3.

Table 7 also shows that our estimate of the average reform effect is not largely driven by comparisons in which earlier adopters of central exams are effectively serving as controls for late adopters of central exams.[27] Importantly, all of these 3 compo-

[24]Note, that in models without country-specific linear trends the larger estimated effects on numeracy and literacy (not reported in Table 5) remain significant even with standard errors clustered at the country level.

[25]We implement the decomposition using the STATA software package bacondecomp provided by Goodman-Bacon et al. (2019). The sample of countries that allows us to implement the decomposition excludes the 3 reform countries with federal education systems and non-binary treatment status, the 2 countries that abolished central exams, Greece and Sweden, as well as two countries with observation gaps in graduation cohorts, Austria and New Zealand.

[26]While in the Goodman-Bacon decomposition all weights are by construction positive, De Chaisemartin and d'Haultfoeuille (2020) show, using a more detailed decomposition, that negative weights are possible and may cause a negative estimate even if all the average treatment effects are positive. Reassuringly, using De Chaisemartin and d'Haultfoeuille (2020)'s twowayfeweights STATA command, we find in our application that only 16 percent of the weights are negative.

[27]The comparisons of later- to earlier-treated groups (the timing groups) are positive on average with weights

nents indicate a positive average effect and comparisons of timing groups is accounting for substantially less of the total effects as opposed to the average treated/untreated estimates.

Finally, we also apply the estimator of average treatment effects in DD setups with multiple time periods and variation in treatment timing developed in Callaway and Sant'Anna (2021) to our data.[28] Results are reported in Table 8. Callaway and Sant'Anna (2021) suggest to exclude always-treated units in the analysis. Thus, we report in the first row of Table 8 estimates of our baseline twoway-fixed effects specifications for an estimation sample of only 16 countries. In this sample, all estimates of the effect on adults are significant for both levels of clustering, but interestingly estimated coefficients now tend to be slightly larger when country-specific linear trend are included in the twoway-fixed effect model.

Reassuringly, estimates based on Callaway and Sant'Anna (2021)'s doubly robust estimator are very similar to the estimated effects in the twoway-fixed effect model. With standard errors clustered at the country level, only the effect on problem-solving skills is statistically significant. However, using wild bootstrapped standard errors, as suggested by Callaway and Sant'Anna (2021), effects on all adult skills are statistically significant.

6.3 Permutation test

To check whether our estimates are driven by an accidental correlation between changes in exam regimes and other unobserved discrete changes that impact skill development, we conduct a placebo analysis in the spirit of the placebo and permutation tests conducted in Chetty et al. (2009).[29] Our placebo check replaces the actual year of the reform with a randomly chosen year between 1960 and 2012 for each of the 13 de facto regime-changing countries. We repeat this randomization procedure 2,000 times, estimating the reform effect on adult skills based on our baseline specification for each of those 2,000 iterations.[30]

---

between 33 and 37 percent. The average treated/untreated estimates are also all positive. Together these two components account for 66 and 62 percent of the total effect. Interestingly, the comparisons of the always vs. timing groups are always significantly higher than the comparisons of never vs. timing groups. This indicates a some heterogeneity in the 2×2 DD estimates, which is also visible in Figure A2.

[28] We implement Callaway and Sant'Anna (2021)'s doubly robust estimator using the STATA command `csdid` (Rios-Avila et al. 2021) with stabilized inverse probability weighting.

[29] Chetty et al. (2009) use the test to check for robustness in their estimates of posting-tax on demand effect. They estimate a placebo effect immediately before and immediately after the posting-tax experiment in order to check for unusual patterns in demand around the timing of the experiment.

[30] Note that the direction of the placebo policy change is assumed to be identical to the actual direction of the reform, i.e., introduction or discontinuation of central exams. For Greece, which experienced two

Figure 2 displays the empirical cumulative distribution of central exam effect estimates from the 2,000 placebo randomizations. The results reveal that more than 95 percent of the placebo estimates are well below our baseline estimate. In most cases, estimates are close to zero and insignificant. Actually, the probability that the baseline effect of 0.068 would appear by an accidental correlation is 0.03. Thus, this exercise reveals that the potential for discrete changes in unobserved impacts on skills that could generate effect sizes comparable to our baseline estimate is very low in our data.

6.4 Exploiting within-country variation

For the German subsample, PIAAC provides identifiers for the respondents' federal state of residence, which allows us to exploit the variation in exam regimes across federal states within a single country. The results should be directly comparable to the findings of Backes-Gellner and Veen (2008) and Piopiunik et al. (2013).

Table 9 presents the results of this analysis. The set of controls used for the skills outcomes (columns 1 to 3) is identical to Table 3, and the set of controls used for labor market outcomes (columns 4 to 6) is identical to Table 4's. Given that we cannot use specification with state (*Länder*) fixed effects, the estimate for numeracy in Table 9 corresponds to the international estimate in column 2 of our baseline table (Table 3).

For numeracy and literacy skills, we find a significant positive association with central exams. The estimates for problem-solving skills are also positive and economically important, but only weakly significant. Our estimated effect of 0.11 for numeracy is fairly close to the preferred effect of central exams to the TIMSS math score, namely 0.13 of national standard deviation, as evidenced by Jürges et al. (2005).

As for labor market outcomes, we do not find a significant link between central exams and earnings, but we do find a positive and highly significant relation with employment probabilities. This pattern is broadly in line with the estimation results of Backes-Gellner and Veen (2008) and Piopiunik et al. (2013), which exploit the same variation but use data from the German Socio-Economic Panel (GSOEP) for their analysis.

While these associations do not necessarily warrant a causal interpretation, overall they support the interpretation of our baseline estimates. In particular, they increase confidence in the validity of our analysis of the labor market effects of central exams in

changes in exam regimes between 1960 and 2012, the randomization assumes that the first randomized year of change must be smaller than the second. For simplicity, in this analysis we dropped the three countries with federal education systems. The data for the 14 countries that didn't change their central exam regime are not manipulated in this placebo analysis.

PIAAC, as we find a pattern of results that is similar to the existing evidence on the same association based on an independent data source.

6.5 Other education reforms

Another concern is that education reforms often come in packages that not only target a specific element (such as introducing central exams) but also impact multiple levels of an education system. Thus, the effect that we are capturing in our analysis might not be attributable solely to changes in the central versus local nature of exit exams if other elements of the education system systematically covary with exit exam systems.

We did further investigate the reforms to the exam systems we have identified, and checked what other features of the education systems might have changed simultaneously. We find that changes to exam systems are indeed, in some cases, embedded in a larger set of reforms that typically share a specific educational philosophy or are aimed at achieving common objectives. We see, however, no clear pattern suggesting that there is indeed a common set of other institutional changes that typically accompanies changes in exit exams in the countries where we can study this.[31]

Overall, our investigation of other reforms does not suggest that there actually was a set of other institutional changes that systematically accompanied the changes in the exam regimes we are studying. The only shared element of all these reforms is that they changed the central versus local nature of exit exams at the end of high school. However, we cannot exclude the possibility that each single estimated reform effect of these 39 reforms, that drive our variation in central exams over time, is also partly capturing other reform effects. Thus, a more careful interpretation of our results is that educational reform packages that include strengthening the central nature of exit exams as one element do lead to higher adult skills on average.

7 Heterogeneous effects

In this section we explore potential effect heterogeneities. We are particularly interested in understanding whether central exam effects vary across different subgroups of the

[31] To provide just two examples, we elaborate further on the cases of Sweden and Greece. Sweden abolished central exams in a comprehensive upper-secondary education reform in 1968. With this reform, the former division of academic and vocational post-compulsory education was rescinded and a common curriculum was implemented for all school tracks. After the 1968 reform, substantial effort was devoted to modernizing vocational education at the upper-secondary level and preparing the work force for the economic challenges of the time (Lundahl et al., 2010; Garrouste, 2010). In Greece, the 1983 introduction of central exams was also surrounded by other education reforms, including reforms to teacher recruitment and training in 1982 and reform in 1985 targeted at the structure and operation of primary and secondary education, and training standards for teachers (Garrouste, 2010).

population and other features of education systems, whether reforms introducing or abolishing central exams have immediate effects on student performance, and whether the effect of central exams on individuals' skills tends to decay over time.

7.1 Effect heterogeneity across population groups and education systems

Our main results show that central exams affect adult skills and labor market outcomes on average. But this may conceal a substantial effect heterogeneity. Guided by evidence on important heterogeneities of central exams in a cross-sectional analysis of student achievement presented in Woessmann (2005), we study whether central exam effects on adult outcomes also differ by gender, migration background, parental education, and school autonomy.

To this end, we estimate variants of Equation 1 including interactions with the individual characteristic indicated in each panel of Table 10. Panel A reports results on potential heterogeneities of central exam effects by three categories of educational background parents. In Panel B we report estimates of an interaction between cental exam and second generation migrants. And Panel C investigates potential effect heterogeneities by gender.

Overall, the estimates in Table 10 suggest that all groups benefit from central exams in terms of higher adult skills. In terms of effect heterogeneities by social background, we find that effects on skills, if anything, may be slightly larger for students from more educated households. Interestingly, the estimates of the effects on earnings point in a different direction. While the effects are not significantly different from zero for individuals from more educated households, individuals whose parents lack a secondary degree actually seem to also benefit from central exams in terms of their later earnings (see column (4) of Panel A). With respect to migration background, results in Panel B show that second-generation migrants benefit slightly more from central exams in terms of their skill development, but differential effects on labor market outcomes do not exist. Panel C suggests that central exam effects on adult skills are largely independent of gender, with some evidence on interactions effects in terms of numeracy skills. We find, however, positive effects on earnings for females.

These findings contribute to the discussion of whether school accountability policies effectively decrease or increase gaps in student outcomes for disadvantaged groups. In fact, it has been shown, in the context of minimum competency exit exams in the United States, that the introduction of such exams increase dropout rates of minority and disadvantaged students. In the context of the literature on the international variation in central exams, Woessmann (2005), however, reports that immigrant students or students from a less-educated family background even benefit more from central exams in terms of student achievement. Our results are in line with these findings.

The most interesting finding of Woessmann (2005) is, however, that central exams have complementary effects to school autonomy. Unfortunately, information on school autonomy and the development thereof over time is not available for our extensive study period from 1960 to 2015. We can, however, operationalize differences in school autonomy based on a time-invariant classification of countries by differences in school autonomy in recent years. For this exercise, we can use measures of school autonomy presented in Hanushek et al. (2013).[32] For simplicity, we use this measure of cross-country differences in school autonomy at the beginning of this century to categorize countries into two groups, low and high autonomy countries, and assume that this classification is a reasonable approximation of longer-run differences in school autonomy over our entire study period.[33]

Table 11 shows results of estimating Equation 1 separately for low and high autonomy countries. In line with Woessmann (2005), we also find evidence for a complementarity between central exams and school autonomy. The central exam effect on adult skills is only significant in high autonomy settings. In these settings we also find that central exams have positive effects on earnings and college graduation rates. This indicates a substantial heterogeneity in the effect of central exams on adult outcomes depending on the level of school autonomy in a country, with the effect being stronger for more autonomous schools. This finding is in line with predictions of a principal-agent model in which central exams act as a monitoring device and in which schools have a significant local knowledge advantage, but also face substantial incentives for local opportunistic behavior when determining course content.

[32] In particular, we use the index for academic-content autonomy of schools of Hanushek et al. (2013) averaged over the years 2000 and 2009. Based on this measure we categorize countries in our study into low and high autonomy countries according to the mean of the averaged index of all countries included in Hanushek et al. (2013)

[33] Countries in the low autonomy group are: Austria, Belgium, Canada, Chile, Germany, Greece, Norway, Slovak Republic, Turkey, and Spain. Countries in the high autonomy group are: Czech Republic, Denmark, Finland, Israel, Ireland, Italy, Japan, Korea, New Zealand, Netherlands, Poland, Sweden, United States, and United Kingdom. For 6 countries of our full sample no information on school autonomy is available in Hanushek et al. (2013).

7.2 Time to effect

The introduction of central exams might not be effective from the onset.[34] On the one hand, this could be the case if teachers and schools initially face problems with implementation or adjustment. On the other hand, it might take some time for the accountability mechanism set in motion by the introduction of central exams to unfold its full impact. If centralized exams indeed make differences in school quality more salient, it might be a few years until parents, teachers, and schools fully react to the incentives set by this new environment and adjust their behavior in ways that will ultimately lead to more effective learning and teaching.

We test for this "time to effect" assertion by estimating variants of Equation 1, which allow for heterogenous impacts of central exams by years of existence of central exams. In particular, we interact the central exam dummy with two dummies indicating whether central exams exist in a country for more than two and five years, respectively.[35]

Table 12 shows the results of this analysis. Our findings suggest that the impact of the introduction of central exams only really unfolds with a delay of about two years. The estimates in columns 2, 4, and 6 show that the central exam effect becomes statistically significant two years after the policy change. Thereafter, the effect of central exams might still increase further (in particular in the case of literacy and problem-solving skills), but any additional average central exam effect five years after the reform is not significantly different from the average effect after two years. However, given our limited statistical power, it is difficult to capture the dynamics in a very nuanced way. For the sample of 16 countries, the event study analysis presented in Figure A1 at least provides some graphical evidence that treatment effects might evolve even more dynamically, with treatment effects becoming larger the longer the central exams are in place in a country.

7.3 Fade-out of central exam effects

Comparing our results with findings of the existing literature on student skills can be informative about the extent of fade-out in the effects of central exams. However, differences in sample populations potentially blur this comparison. We address this concern at least partly by investigating simple correlations between central exams and

[34]In a different policy context, for example, Hanushek and Raymond (2005) show gradual growth of the acceptance of the National Assessment of Educational Progress criteria across U.S. states over the 1993-2002 period.

[35]For presentation reasons, we chose a model with just two interactions at two and five years after the policy change. A more flexible model with interactions for each year reveals that this split roughly captures the overall pattern of effect heterogeneities quite well.

the skills of both students and adults for samples drawn from the same population. Specifically, PISA 2000 surveyed a representative sample of 15-year-old students in 2000. Thus, the PISA sample should be sampled from about the same population as the subsample of PIAAC participants who were 15 years old in 2000. Similarly, we construct a correspondence data set for the TIMSS Repeat (1998/1999) data.

We report the results of this exercise in Table 13. All estimates stem from simple regressions of skills on a central exam dummy that differs across countries. Unfortunately, there is only a limited set of countries that participated in both PIAAC and PISA 2000 (columns 1 and 2) and countries that participated in both PIAAC and TIMSS Repeat (columns 3 and 4). The estimates are all insignificant, but the estimated coefficients in the student data are large and roughly comparable in size to the analogous estimates reported in studies that use the PIAAC 2000 and TIMSS Repeat data to study the central exam effect on student skills (for example, Woessmann, 2002; Fuchs and Woessmann, 2007; Jürges et al., 2005). Most notably, the estimated coefficients on the central exam dummy are one third to one half lower if the dependent variable is adult skills rather than student skills. While the results of this purely descriptive comparison must be interpreted with caution, the decrease in estimated coefficients would be in line with central exam effects that fade out over time.

8 Conclusion

This paper is the first to investigate the relationship between the type of exit examination at the end of secondary school and the cognitive skills and labor market outcomes of adults, using a panel framework. We find that central exams increase the cognitive skills of adults on average. But compared to the results of studies that examine the effects on student skills, our estimates of the long-run effects on adults skills are small. We find effect sizes as low as 7 percent of a standard deviation. This finding might be the consequence of some fade-out in central exam effects over time.

In terms of labor market outcomes, we do not detect any significant effects. In particular, our statistical power allows us to rule out effects on earnings above 4.6 percent. This finding is consistent with recent evidence on the average wage return to increases in adult skills of about 7 percent of a standard deviation, if we assume that central exams affect earnings only through their impact on skill production.

In line with previous evidence on heterogeneities of central exams on student achievement presented in Woessmann (2005), we find effect heterogeneity across countries and population groups. In particular, we find evidence for a complementarity between central exams and school autonomy, evidence for stronger effects on skills for second-generation migrants and some evidence for positive earnings effects for females and individuals with low parental education background.

These findings contribute to the growing literature on accountability systems in schools in general, and the literature on central exams in particular. Overall, our results corroborate the notion of the positive effects of central exams on skill development.

References

Abadie, A., S. Athey, G. W. Imbens, and J. Wooldridge (2022). When should you adjust standard errors for clustering? Updated version of NBER working paper 24003. https://economics.mit.edu/files/13927.

Angrist, J. D. and J.-S. Pischke (2015). *Mastering'metrics: The path from cause to effect*. Princeton University Press.

Backes-Gellner, U. and S. Veen (2008). The consequences of central examinations on educational quality standards and labour market outcomes. *Oxford Review of Education 34*(5), 569–588.

Baker, O. and K. Lang (2013). The effect of high school exit exams on graduation, employment, wages and incarceration. Working Paper 19182, National Bureau of Economic Research.

Becker, W. E. and S. Rosen (1992). The learning effect of assessment and evaluation in high school. *Economics of Education Review 11*(2), 107–118.

Bergbauer, A. B., E. A. Hanushek, and L. Woessmann (2021). Testing. *Journal of Human Ressources*, forthcoming, doi: 10.3368/jhr.0520–10886R1.

Bethell, G. and A. Zabulionis (2000). *Examination Reform in Lithuania: Background, Strategies and Achievements*. National Examination Centre.

Betts, J. R. (1998). The impact of educational standards on the level and distribution of earnings. *American Economic Review 88*(1), 266–275.

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 economic efficiency? *Swedish Economic Policy Review 6*(2), 349–401.

Bishop, J. H. (1999b). Nerd harassment, incentives, school priorities and learning. In S. E. Mayer and P. E. Peterson (Eds.), *Earning and learning: How schools matter*. Washington, DC. 231–280.

Bishop, J. H. (2006). Drinking from the fountain of knowledge: Student incentive to study and learn - externalities, information problems, and peer pressure. In E. A.

Hanushek and F. Welch (Eds.), *Handbook of the Economics of Education*, Volume 2. Amsterdam: North-Holland. 909–944.

Bishop, J. H. and F. Mane (2001). The impacts of minimum competency exam graduation requirements on high school graduation, college attendance and early labor market success. *Labour Economics 8*(2), 203–222.

Bishop, J. H., J. Y. Moriarty, and F. Mane (2000). Diplomas for learning, not seat time: The impacts of new york regents examinations. *Economics of Education Review 19*(4), 333–349.

Bishop, J. H. and L. Woessmann (2004). Institutional effects in a simple model of educational production. *Education Economics 12*(1), 17–38.

Boezerooy, P. and J. Huisman (2000). From secondary to tertiary education in the Netherlands. *Journal of Institutional Research 9*(1), 54–64.

Callaway, B. and P. H. Sant'Anna (2021). Difference-in-differences with multiple time periods. *Journal of Econometrics 225*(2), 200–230.

Caves, K. and S. Balestra (2018). The impact of high school exit exams on graduation rates and achievement. *Journal of Educational Research 111*(2), 186–200.

Chang, B.-M. (2009). Korea's English education policy innovations to lead the nation into the globalized world. *Journal of Pan-Pacific Association of Applied Linguistics 13*(1), 83–97.

Chetty, R., A. Looney, and K. Kroft (2009). Salience and taxation: Theory and evidence. *American Economic Review 99*(4), 1145–1177.

Clarke, D. and K. Tapia-Schythe (2021). Implementing the panel event study. *The Stata Journal 21*(4), 853–884.

Costrell, R. M. (1997). Can centralized educational standards raise welfare? *Journal of Public Economics 65*(3), 271–293.

Cullen, J. B. and R. Reback (2006). Tinkering toward accolades: School gaming under a performance accountability system. In T. Gronberg and D. Jansen (Eds.), *Improving School Accountability: Check-Ups or Choice, Advances in Applied Microeconomics 14*. Amsterdam: Elsevier Science. 1–34.

De Chaisemartin, C. and X. d'Haultfoeuille (2020). Two-way fixed effects estimators with heterogeneous treatment effects. *American Economic Review 110*(9), 2964–96.

Dee, T. S. and B. A. Jacob (2007). Do high school exit exams influence educational attainment or labor market performance? In A. Gamoran (Ed.), *Standards-based reform and the poverty gap: Lessons for no child left behind*. Washington, D.C.: Brookings University Press.

Eckstein, M. A. and H. J. Noah (1989). Forms and functions of secondary-school-leaving examinations. *Comparative Education Review 33*(3), 295–316.

Entrich, S. R. (2015). The decision for shadow education in japan: Students' choice or parents' pressure? *Social Science Japan Journal 18*(2), 193–216.

Figlio, D. and S. Loeb (2011). School accountability. In E. A. Hanushek, S. Machin, and L. Woessmann (Eds.), *Handbook of the Economics of Education*, Volume 3. Amsterdam: North-Holland. 383–421.

Figlio, D. N. and L. S. Getzler (2006). Accountability, ability and disability: Gaming the system? In T. Gronberg and D. Jansen (Eds.), *Improving School Accountability: Check-Ups or Choice, Advances in Applied Microeconomics 14*. Amsterdam: Elsevier Science. 35–49.

Fuchs, T. and L. Woessmann (2007). What accounts for international differences in student performance? A re-examination using PISA data. *Empirical Economics 32*(2-3), 433–464.

Garrouste, C. (2010). *100 years of educational reforms in Europe: A contextual database*. EUR 24487 EN, Luxembourg: Publications Office of the European Union.

Goodman-Bacon, A. (2021). Difference-in-differences with variation in treatment timing. *Journal of Econometrics 225*(2), 254–277.

Goodman-Bacon, A., T. Goldring, and A. Nichols (2019). bacondecomp: Stata module for decomposing difference-in-differences estimation with variation in treatment timing. Stata module. ideas.repec.org/c/boc/bocode/s458676.html.

Goulas, S. and R. Megalokonomou (2021). Knowing who you actually are: The effect of feedback on short- and longer-term outcomes. *Journal of Economic Behavior and Organization 183*, 589–615.

Greger, D. and E. Kifer (2012). Lost in translation: Ever changing and competing purposes for national examinations in the Czech Republic. *Journal of Pedagogy / Pedagogický Časopis 3*(1), 43–81.

Hall, C. (2000). National certificate of educational achievement: Issues of reliability, validity and manageability. *New Zealand Annual Review of Education 9*, 173–196.

Hanushek, E. A., S. Link, and L. Woessmann (2013). Does school autonomy make sense everywhere? Panel estimates from PISA. *Journal of Development Economics 104*, 212–232.

Hanushek, E. A. and M. E. Raymond (2005). Does school accountability lead to improved student performance? *Journal of Policy Analysis and Management 24*(2), 297–327.

Hanushek, E. A., G. Schwerdt, S. Wiederhold, and L. Woessmann (2015). Returns to skills around the world: Evidence from PIAAC. *European Economic Review 73*(1), 103–130.

Hanushek, E. A., G. Schwerdt, S. Wiederhold, and L. Woessmann (2017). Coping with change: International differences in the returns to skills. *Economics Letters 153*, 15–19.

Hanushek, E. A. and L. Woessmann (2011). The economics of international differences in educational achievement. In E. A. Hanushek, S. Machin, and L. Woessmann (Eds.), *Handbook of the Economics of Education*, Volume 3, pp. 89–200. Amsterdam: North Holland.

Hörner, W., H. Döbert, B. von Kopp, and W. Mitter (2007). *The education systems of Europe*. Springer. doi: 10.1007/1-4020-4874-2.

Hvidman, U. and H. H. Sievertsen (2021). High-stakes grades and student behavior. *Journal of Human Resources 56*(3), 821–849.

Ilc, G., V. R. Gabrovec, and A. Stopar (2014). Relating the Slovenian secondary school English language national examinations to the CEFR: Findings and implications. *Linguistica 54*(1), 293–308.

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

Jones, R. S. (2013). Education reform in Korea. Working Paper 1067, OECD Economics Department. doi: 10.1787/5k43nxs1t9vh-en.

Jürges, H., W. F. Richter, and K. Schneider (2005). Teacher quality and incentives: Theoretical and empirical effects of standards on teacher quality. *Finanz Archiv: Public Finance Analysis 61*(3), 298–326.

Jürges, H. and K. Schneider (2010). Central exit examinations increase performance... but take the fun out of mathematics. *Journal of Population Economics 23*(2), 497–517.

Jürges, H., K. Schneider, and F. Büchel (2005). The effect of central exit examinations on student achievement: Quasi-experimental evidence from TIMSS Germany. *Journal of the European Economic Association 3*(5), 1134–1155.

Jürges, H., K. Schneider, M. Senkbeil, and C. H. Carstensen (2012). Assessment drives learning: The effect of central exit exams on curricular knowledge and mathematical literacy. *Economics of Education Review 31*(1), 56–65.

Klein, E. D., S. M. Kühn, I. v. Ackeren, and R. Block (2009). Wie zentral sind zentrale Prüfungen? Abschlussprüfungen am Ende der Sekundarstufe II im nationalen und internationalen Vergleich. *Zeitschrift für Pädagogik 55*(4), 596–621.

Klinger, D. (2008). The evolving culture of large-scale assessments in Canadian education. *Canadian Journal of Educational Administration and Policy 76*, 1–34.

Koretz, D. M. (2002). Limitations in the use of achievement tests as measures of educators' productivity. *Journal of Human Resources 37*(4), 752–777.

Lamprianou, I. (2012). Unintended consequences of forced policy-making in high stakes examinations: The case of the Republic of Cyprus. *Assessment in Education: Principles, Policy & Practice 19*(1), 27–44.

Lavy, V. (2008). Does raising the principal's wage improve the school's outcomes? Quasi-experimental evidence from an unusual policy experiment in Israel. *Scandinavian Journal of Economics 110*(4), 639–662.

Lim, E. P. Y. and A. Tan (1999). Educational assessment in Singapore. *Assessment in Education: Principles, Policy & Practice 6*(3), 391–404.

Lüdemann, E. (2011). Intended and unintended short-run effects of the introduction of central exit exams. In E. Lüdemann (Ed.), *Schooling and the formation of cognitive and non-cognitive outcomes*, Ifo Beiträge zur Wirtschaftsforschung 39. Ifo Institut.

Lundahl, L., I. E. Arreman, U. Lundström, and L. Rönnberg (2010). Setting things right? Swedish upper secondary school reform in a 40-year perspective. *European Journal of Education 45*(1), 46–59.

Metzger, C., R. Maynard, J. Vultaggio, T. Daizen, S. Promboon, K. Y. Ip, and S.-Y. Park (2010). A comparative perspective on the secondary and post-secondary education systems in six nations: Hong Kong, Japan, Switzerland, South Korea, Thailand and the United States. *Procedia-Social and Behavioral Sciences 2*(2), 1511–1519.

Nardi, E. (2001). The transition from school to university in Italy: Examination reform and outstanding issues. *Assessment in Education: Principles, Policy & Practice 8*(3), 339–351.

OECD (2002). Reviews of national policies for education: Lithuania. doi: 10.1787/9789264193475-en.

OECD (2010). OECD economic surveys: Chile. doi: 10.1787/19990847.

OECD (2013a). OECD skills outlook 2013: First results from the survey of adult skills. doi: 10.1787/9789264204256-en.

OECD (2013b). Technical report of the survey of adult skills (PIAAC). Available at www.oecd.org/site/piaac/_Technical%20Report_17OCT13.pdf.

OECD (2016). Skills matter: Further results from the survey of adult skills. doi: 10.1787/9789264258051-en.

Piopiunik, M., G. Schwerdt, and L. Woessmann (2013). Central school exit exams and labor-market outcomes. *European Journal of Political Economy 31*(1), 93–108.

Pischke, J.-S. (2007). The impact of length of the school year on student performance and earnings: Evidence from the German short school years. *Economic Journal 117*(523), 1216–1242.

Pischke, J.-S. and T. von Wachter (2008). Zero returns to compulsory schooling in Germany: Evidence and interpretation. *The Review of Economics and Statistics 90*(3), 592–598.

Polydorides, G. (1986). The determinants of educational achievement at the end of secondary schooling: The case of Greece. Conference Paper ED 270048, ERIC: Institute of Education Sciences.

Popham, W. J. (2001). Teaching to the test? *Educational Leadership 58*(6), 16–21.

Rios-Avila, F., P. H. Sant'Anna, and B. Callaway (2021). CSDID: Stata module for the estimation of Difference-in-Difference models with multiple time periods. Statistical Software Components, Boston College Department of Economics.

Schwerdt, G. and L. Woessmann (2017). The information value of central school exams. *Economics of Education Review 56*, 65–79.

Stephens, Melvin, J. and D.-Y. Yang (2014). Compulsory education and the benefits of schooling. *American Economic Review 104*(6), 1777–1792.

Strachan, J. (2002). Assessment in change: some reflections on the local and international background to the national certificate of educational achievement (ncea). *New Zealand Annual Review of Education 11*(11), 245–262.

Sun, L. and S. Abraham (2021). Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *Journal of Econometrics 225*(2), 175–199.

Syrquin, A. (1997). *The EFL final high school examinations in Israel (bagrut): Historical analysis.* Ph. D. thesis, American University, Washington.

Volante, L. (2004). Teaching to the test: What every educator and policy-maker should know. *Canadian Journal of Educational Administration and Policy* (35), 1–6.

Volante, L. (2006). An alternative vision for large–scale assessment in canada. *Journal of Teaching and Learning 4*(1), 1–14.

Watanabe, Y. (2013). The national center test for university admissions. *Language Testing 30*(4), 565–573.

Woessmann, L. (2002). Central exams improve educational performance: International evidence. Kiel Discussion Paper 397, Kiel Institute for the World Economy.

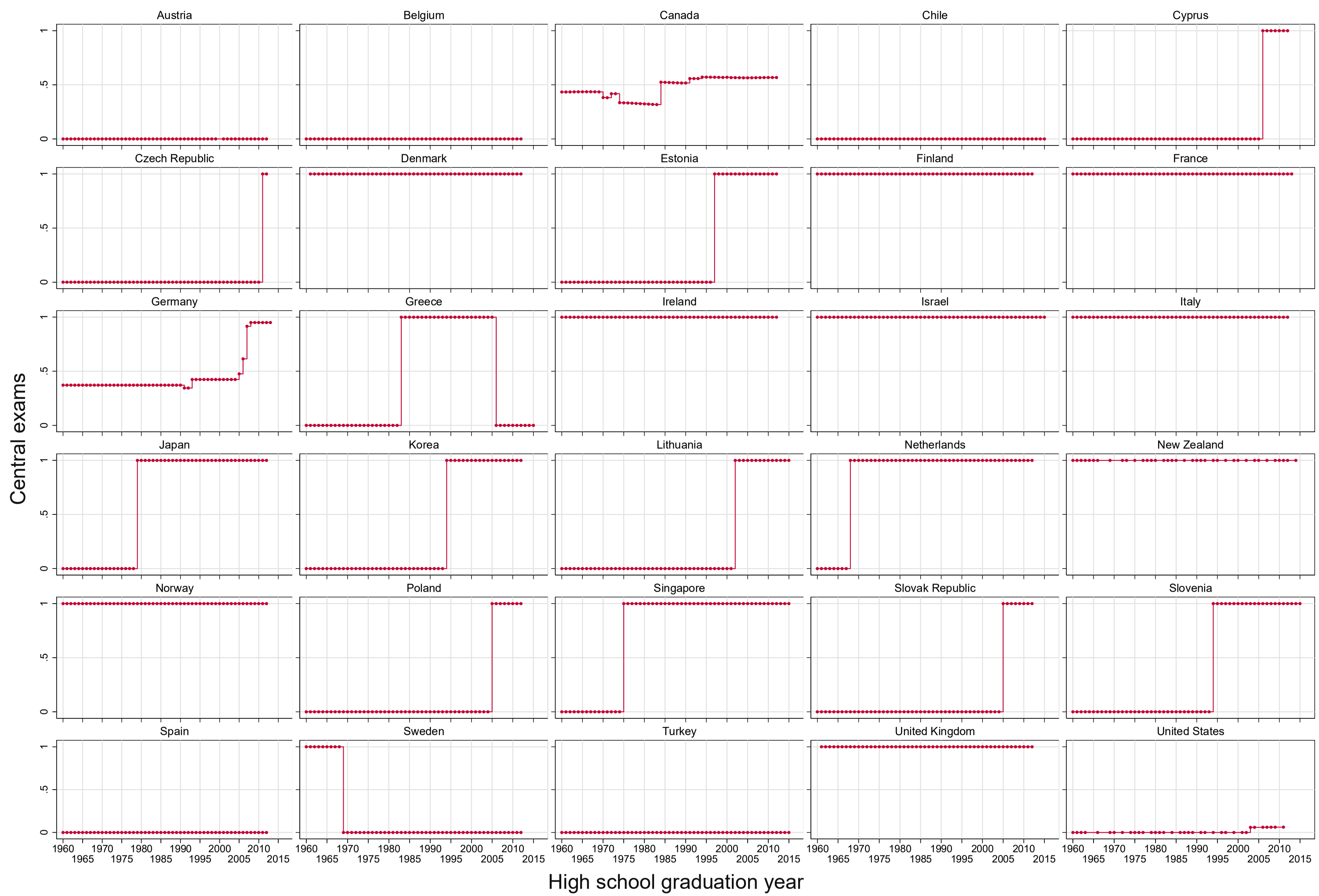
Woessmann, L. (2003). Schooling resources, educational institutions and student performance: The international evidence. *Oxford Bulletin of Economics and Statistics 65*(2), 117–170.

Woessmann, L. (2005). The effect heterogeneity of central examinations: Evidence from TIMSS, TIMSS-Repeat and PISA. *Education Economics 13*(2), 143–169.

Yemini, M., H. Bar-Nissan, and O. Yardeni (2014). Between "us" and "them": Teachers' perceptions of the national versus international composition of the Israeli history curriculum. *Teaching and Teacher Education 42*(1), 11–22.

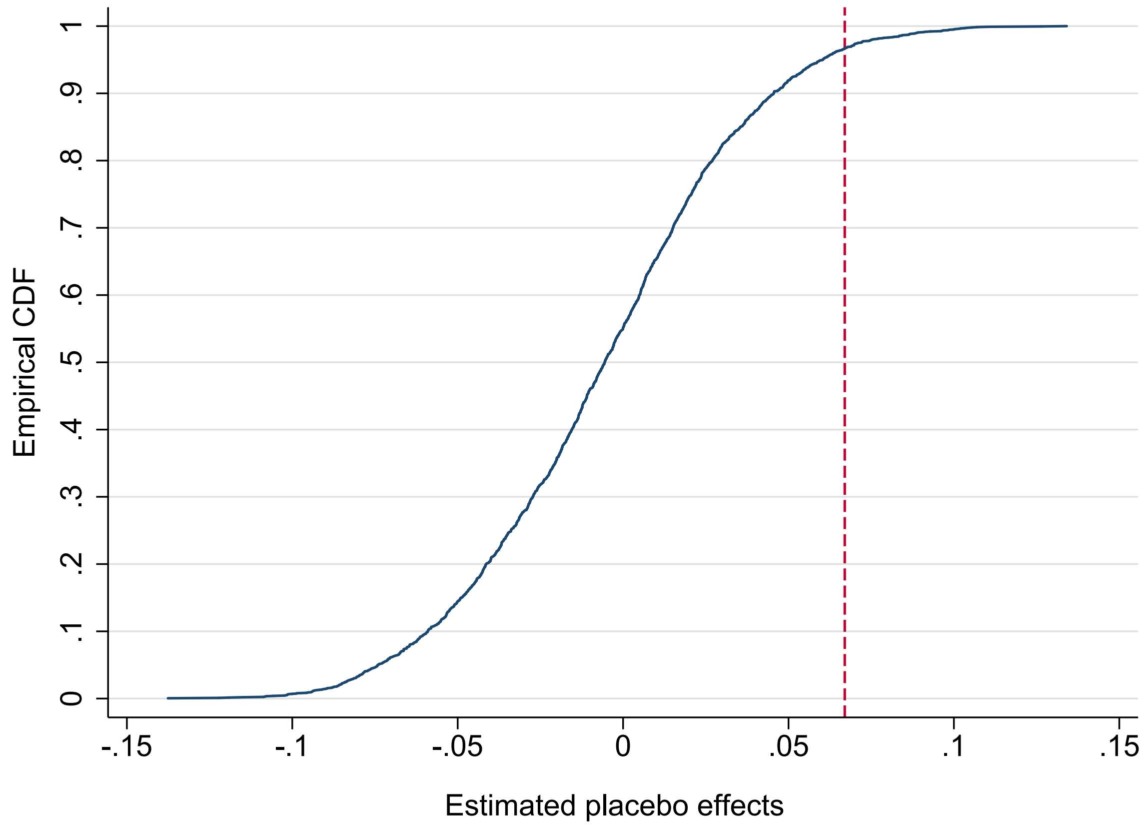
Yildirim, I., T. Ergene, and K. Munir (2007). High rates of depressive symptoms among senior high school students preparing for national university entrance examination in Turkey. *The International Journal on School Disaffection 4*(2), 35–44.

Figure 1: Central exam regimes in PIAAC countries (1960-2015)



Note: Figure displays the evolution of the exposure to central exams in 30 countries by high school graduation cohorts. In the three countries with federal education systems, namely Canada, Germany, and United States, the exposure to central exams is measured as the share of the population exposed to central exams.

Figure 2: Distribution of central exam effect estimates from placebo randomizations



Note: The figure displays the empirical distribution of placebo estimates of central exam effects based on 2,000 randomizations of central exam change. The randomizations are based on the 27-countries sample that excludes the three federal education systems, and use the specification (4) of Table 3. The vertical line corresponds to the effect estimate of .067 for the sample without the three federal education systems reported in Table 6.

Table 1: Central exams status across PIAAC countries 1960-2015

Central exam countries - always treated				
Denmark	Finland	France	Ireland	Israel
Italy	New Zealand	Norway	United Kingdom	
Non-central exam countries - never treated				
Austria*	Belgium (Flanders)	Chile	Spain	Turkey
Reform countries - with national education systems				
Cyprus	Czech Republic	Estonia	Greece	Japan
(2006)	(2011)	(1997)	(1983) [2005]	(1979)
Korea	Lithuania	Netherlands	Poland	Singapore
(1994)	(2002)	(1968)	(2005)	(1975)
Slovak Republic	Slovenia	Sweden		
(2005)	(1994)	[1968]		

Notes: Years in round/square brackets indicate the year in which central exams were introduced/abolished. See Table A1 for a detailed list of sources. *Austria introduced central exams, the *Zentralmatura*, for all schools in 2016.

Table 2: Descriptive statistics

	No central exams		Central exams		
	Mean	Std. dev.	Mean	Std. dev.	Mean diff.
	(1)	(2)	(3)	(4)	(5)
Numeracy	.045	(.965)	.234	(.927)	.189
Literacy	.008	(.938)	.244	(.911)	.236
Problem-solving	-.125	(.990)	.139	(.935)	.264
Earnings	2.44	(.754)	2.80	(.637)	.357
Employment	.923		.935		.012
College degree	.347		.436		.089
High school	.885		.891		.006
Age	43.02	(13.89)	36.40	(14.16)	-6.62
Age at graduation	18.49	(3.37)	18.77	(4.23)	.281
Female	.495		.498		.003
Second-gen. migrant	.094		.135		.041
Parent education:					
Primary	.400		.319		-.081
Secondary	.366		.351		-.015
Above secondary	.197		.272		.075
Missing	.037		.058		.021
Observations	75,387		85,965		

Notes: Means and standard deviations by exam regime, for variables used in empirical analysis. Individuals from Germany, Canada, and the USA are included in the central exam group if their central exam factor is > 0.5. Otherwise they are classified as non-central exam. Observations are weighted by adjusted sampling weights that give same weight to each country. Baseline sample: high-school graduates aged 16-65. Numbers of observations are smaller, than those indicated, for the following variables: for earnings, employment, college degree, and high school attainment where we exclude those still studying and those younger than 25; problem-solving skills are not available for Cyprus, France, Italy, and Spain.

Table 3: The effect of central exams on adult skills

	Numeracy skills				Literacy		Prob. solv.	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Central exams	.211***	.170***	.092***	.068***	.075***	.068***	.136***	.092***
	(.019)	(.017)	(.016)	(.021)	(.017)	(.022)	(.018)	(.022)
Female		-.188***	-.188***	-.188***	-.017***	-.017***	-.090***	-.091***
		(.006)	(.006)	(.006)	(.005)	(.005)	(.008)	(.008)
Second-gen. migrant		-.019	-.024**	-.009	-.030***	-.011	-.016	-.008
		(.012)	(.010)	(.011)	(.010)	(.011)	(.013)	(.013)
Parent education:								
Secondary		.275***	.229***	.229***	.231***	.232***	.254***	.248***
		(.009)	(.008)	(.008)	(.008)	(.008)	(.011)	(.011)
Above secondary		.569***	.519***	.520***	.515***	.515***	.568***	.561***
		(.010)	(.010)	(.010)	(.010)	(.009)	(.012)	(.012)
Educ. missing		-.218***	-.204***	-.201***	-.164***	-.161***	-.042**	-.046**
		(.018)	(.015)	(.015)	(.015)	(.015)	(.021)	(.021)
Grad. year FE	yes	yes	yes	yes	yes	yes	yes	yes
Grad. age FE	no	yes	yes	yes	yes	yes	yes	yes
Country FE	no	no	yes	yes	yes	yes	yes	yes
Country-specific linear trend in grad. year	no	no	no	yes	no	yes	no	yes
Countries	30	30	30	30	30	30	26	26
Observations	161,352	161,352	161,352	161,352	161,352	161,352	113,417	113,417
Clusters	1,478	1,478	1,478	1,478	1,478	1,478	1,245	1,245
R^2	.101	.186	.253	.259	.246	.253	.230	.234

Notes: Least squares regressions weighted by adjusted sampling weights that give same weight to each country. Dependent variables are the globally standardized test scores indicated in first row. Sample: high school graduates aged 16-65. Robust standard errors clustered by graduation year × country in parentheses. * p<0.10, ** p<0.05, *** p<0.01

Table 4: The effect of central exams on labor market and education outcomes

	Earnings	Employed	College	High school
	(1)	(2)	(3)	(4)
Central exams	.010	-.005	.025	-.012
	(.019)	(.006)	(.019)	(.015)
Female	-.176***	.004*	.051***	.013***
	(.006)	(.002)	(.005)	(.002)
Second-gen. migrant	.009	-.014***	-.002	.003
	(.009)	(.004)	(.007)	(.004)
Parents education:				
Secondary	.110***	.011***	.095***	.039***
	(.007)	(.003)	(.005)	(.003)
Above secondary	.229***	.013***	.313***	.045***
	(.008)	(.003)	(.007)	(.003)
Educ. missing	-.028*	-.029***	-.039***	-.016*
	(.015)	(.007)	(.009)	(.009)
Grad. year FE	yes	yes	yes	yes
Grad. age FE	yes	yes	yes	yes
Country FE	yes	yes	yes	yes
Country-specific linear trend in grad. year	yes	yes	yes	yes
Countries	30	30	30	30
Observations	70,915	76,025	62,485	76,025
Clusters	1,383	1,392	1,313	1,392
R^2	.457	.087	.323	.547

Notes: Least squares regressions weighted by adjusted sampling weights that give same weight to each country. Dependent variables are log gross hourly wage (column 1), an employment dummy (column 2), a dummy for tertiary education degree (column 3), and a high school attainment dummy (column 4). Sample: individuals aged 25-65 with non-zero earnings (column 1), individuals aged 25-65 currently not in education (columns 2, 3, 4). Robust standard errors clustered by graduation year × country in parentheses. * p<0.10, ** p<0.05, *** p<0.01

Table 5: Sensitivity analysis

	Adult skills			Labor market & education		
	Numeracy (1)	Literacy (2)	Prob. solv. (3)	Earnings (4)	Employed (5)	College (6)
Baseline	.068***	.068***	.092***	.010	-.005	.025
	(.021)	(.022)	(.022)	(.019)	(.006)	(.019)
	[.046]	[.055]	[.042]	[.026]	[.009]	[.025]
No weights	.085***	.093***	.094***	.015	-.003	.036**
	(.019)	(.019)	(.018)	(.017)	(.006)	(.014)
Median grad. age	.078***	.068***	.084***	.012	.006	.015
	(.018)	(.019)	(.020)	(.015)	(.005)	(.016)
Incl. early leavers	.078***	.068***	.084***	.013	.007	.013
	(.022)	(.022)	(.020)	(.015)	(.005)	(.016)

Notes: Each cell reports an estimate from a separate regression weighted by adjusted sampling weights that give same weight to each country, except for the second row, where we forgo using any weight. Only the estimate of the central exam effect based on variants of Equation 1 is reported. Sample of 30 (26) countries (in column 3). Dependent variables are globally standardized test scores in numeracy (column 1), literacy (column 2), problem-solving (column 3), log gross hourly wage (column 4), an employment dummy (column 5), and a dummy for tertiary education degree (column 6). "Incl. early leavers" row additionally includes individuals who, based on the calculatory age of graduation (highest attained degree), is below 15, i.e. earlier than the compulsory leaving age in OECD countries. Robust standard errors clustered by graduation year × country in parentheses. Clustered standard errors by country in square brackets. * p<0.10, ** p<0.05, *** p<0.01 (significance refers to clustered standard errors by graduation year × country)

Table 6: Piecewise deletion of countries

	Numeracy (1)	Earnings (2)		Numeracy (3)	Earnings (4)
Excluding one country					
Austria	.068***	.010	Belgium	.067***	.010
	(.021)	(.019)		(.021)	(.019)
Canada	.069***	.011	Chile	.067***	.010
	(.021)	(.019)		(.021)	(.019)
Cyprus	.062***	.009	Czech Republic	.072***	.011
	(.023)	(.019)		(.022)	(.019)
Denmark	.067***	.010	Estonia	.060***	.012
	(.021)	(.019)		(.022)	(.020)
Finland	.068***	.010	France	.069***	.010
	(.021)	(.019)		(.021)	(.019)
Germany	.069***	.009	Greece	.122***	.033*
	(.021)	(.019)		(.025)	(.020)
Ireland	.068***	.010	Israel	.066***	.010
	(.021)	(.019)		(.021)	(.019)
Italy	.068***	.010	Japan	.071***	-.007
	(.021)	(.019)		(.023)	(.019)
Korea	.072***	.005	Lithuania	.055**	.006
	(.022)	(.020)		(.022)	(.020)
Netherlands	.072***	.015	New Zealand	.068***	.010
	(.022)	(.019)		(.021)	(.019)
Norway	.066***	.010	Poland	.057**	.014
	(.021)	(.019)		(.022)	(.019)
Singapore	.068***	.007	Slovak Republic	.053**	.008
	(.017)	(.019)		(.022)	(.019)
Slovenia	.071***	.013	Spain	.067***	.011
	(.022)	(.020)		(.021)	(.019)
Sweden	.071***	.004	Turkey	.068***	.009
	(.022)	(.020)		(.021)	(.019)
United Kingdom	.069***	.009	United States	.068***	.009
	(.021)	(.019)		(.021)	(.019)
Excluding country groups					
Germany, Poland,	.055***	.011	Canada, Greece,	.135***	.017
and Singapore	(.017)	(.020)	Japan, and U.S.	(.028)	(.021)
Japan and Korea	.076***	-.014	Early or late regime	.078***	.010
	(.024)	(.020)	changing countries	(.024)	(.021)

Notes: Each estimate stems from a separate regression excluding one or more countries. The regressions are weighted by adjusted sampling weights that give same weight to each country. We report here only the estimates of the central exam effect on numeracy (column (1)), which corresponds to the specification of Table 3, column (4), and the effect on earnings (column (2)), which corresponds to the specification of Table 4, column (1). Dependent variable is the globally standardized test scores in numeracy, or log gross hourly wage, respectively. Early or late regime changing countries are Netherlands and Sweden (both 1968), Czech Republic (2011). Robust standard errors clustered by graduation year × country in parentheses.
* p<0.10, ** p<0.05, *** p<0.01

Table 7: Goodman-Bacon decomposition of the average central exam effect on adult skills

	Numeracy (1)	Literacy (2)	Weight (3)	Prob. solv. (4)	Weight (5)
Timing groups (36/28)	0.049	0.074	0.3255	0.116	0.3678
Always vs. timing (9/8)	0.168	0.127	0.4396	0.184	0.4239
Never vs. timing (9/8)	0.026	0.056	0.2119	0.148	0.1957
Always vs. never (1)	0.445	-0.808	0.0007	-0.453	0.0007
Within component (1)	-1.049	-1.105	0.0222	-2.239	0.0120
Total effect	0.072	0.067		0.123	
(s.e.)	(.028)	(0.026)		(0.026)	

Notes: Goodman-Bacon decomposition of the average effects of central exams on adult skills into different types of 2×2 DD estimates. The estimate of the total average effect is reported in the bottom row. Estimates are based on Equation 1 without country-specific linear trends for a reduced sample of countries. Sample of 23 countries. Excluded countries are: Austria, Canada, Germany, Greece, New Zealand, Sweden, and the United States. The table reports average contributions and weights for each type of 2×2 DD estimates. Timing groups refer to 2×2 DD estimates for later- to earlier-treated groups. In total, we have $36 + 2 * 9 + 2 = 56$ distinct 2×2 DD estimates for numeracy and literacy skills and $28 + 2 * 8 + 2 = 46$ distinct 2×2 DD estimates effects for problem solving skills.

Table 8: Average effects based on Callaway and Sant'Anna (2021)'s doubly robust estimator

	Numeracy skills		Literacy		Prob. solv.	
	(1)	(2)	(3)	(4)	(5)	(6)
TWFE (Baseline, 16 countries)	.071	.115	.078	.136	.130	.151
	(.020)	(.025)	(.020)	(.024)	(.021)	(.025)
	[.035]	[.025]	[.030]	[.025]	[.032]	[.044]
Callaway and Sant'Anna (2021)	.097		.111		.158	
	[.118]		[.099]		[.066]	
	{.032}		{.031}		{.038}	
Country-specific linear trend	no	yes	no	yes	no	yes

Notes: Effects of central exams on adult skills based on a sample of 16 countries: 5 never treated and 11 reform countries with Sweden and Greece excluded (see Table 1). The first row reports TWFE estimates without and with country-specific linear time trend for each adult skill based on the reduced sample of 16 countries. The second row reports estimates based on Callaway and Sant'Anna (2021)'s doubly robust estimator of the central exam effect using stabilized inverse probability weighting applying `csdid` STATA command (Rios-Avila et al. 2021). To construct bootstrapped standard errors the command uses a multiplicative WildBootstrap procedure with 999 repetitions. Robust standard errors clustered by graduation year × country in parentheses. Clustered standard errors by country in square brackets. Wild bootstrapped standard errors in curly brackets.

Table 9: Central exams and adult outcomes in the German subsample

	Adult skills			Labor market & education		
	Numeracy (1)	Literacy (2)	Prob. solv. (3)	Earnings (4)	Employed (5)	College (6)
Central exams	.112***	.100***	.056*	−.005	.053***	.017
	(.030)	(.032)	(.033)	(.023)	(.016)	(.013)
Backgr. char.	yes	yes	yes	yes	yes	yes
Grad. age FE	yes	yes	yes	yes	yes	yes
Grad. year FE	yes	yes	yes	yes	yes	yes
Observations	4,086	4,086	3,646	2,098	2,949	3,250
Clusters	570	570	555	449	499	543
R^2	.288	.279	.263	.255	.168	.506

Notes: Dependent variables: for numeracy, literacy and problem solving skills we use standardized scores in the German sub-sample. Earnings are log gross hourly wage, employed is a binary employment indicator, college is a dummy indicating that individuals completed any college. The college estimation excludes those currently studying, the earnings and employed estimations additionally exclude those younger than 25. The set of controls used for the skill outcomes (column 1-3) is identical to our preferred specification of Table 3 (column 4), the set of controls used for labor market outcomes (column 4-6) is identical to Table 4. Estimations are weighted by sampling weights. Robust standard errors (in parentheses) clustered at graduation year × federal state. * p<0.10, ** p<0.05, *** p<0.01

Table 10: Effect heterogeneity across population groups

Outcome	Adults skills			Labor market & education		
	Numeracy	Literacy	Prob. solv.	Earnings	Employed	College
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Interactions with parental education						
Central exams	.060**	.045*	.114***	.046**	.002	.029
	(.025)	(.025)	(.026)	(.020)	(.006)	(.021)
Secondary	.008	.029**	-.034*	-.048***	-.011*	.001
× Central exams	(.016)	(.015)	(.020)	(.013)	(.006)	(.011)
Tertiary	.035*	.059***	-.029	-.089***	-.014**	-.021
× Central exams	(.020)	(.019)	(.024)	(.017)	(.007)	(.013)
Panel B: Interactions with migration background						
Central exams	.063***	.064***	.085***	.006	-.004	.027
	(.021)	(.022)	(.022)	(.019)	(.006)	(.019)
Second-gen. migrant	.047*	.036	.075***	.027	-.004	-.012
× Central exams	(.025)	(.024)	(.029)	(.020)	(.008)	(.018)
Panel C: Interactions with gender						
Central exams	.084***	.073***	.095***	-.010	-.007	.021
	(.022)	(.023)	(.024)	(.020)	(.006)	(.020)
Female	-.032***	-.009	-.006	.044***	.005	.010
× Central exams	(.012)	(.011)	(.016)	(.013)	(.005)	(.010)

Notes: Each panel reports least squares regressions weighted by adjusted sampling weights that give same weight to each country. Only estimates of the central exam effects based on variants of Equation 1 including interactions the individual characteristics indicated in each panel are reported. Dependent variables are globally standardized test scores in numeracy (column 1), literacy (column 2), problem-solving (column 3), log gross hourly wage (column 4), an employment dummy (column 5), and a dummy for tertiary education degree (column 6). Robust standard errors clustered by graduation year × country in parentheses. * p<0.10, ** p<0.05, *** p<0.01

Table 11: Complementarity between central exam and school autonomy

	Adults skills			Labor market & education		
	Numeracy	Literacy	Prob. solv.	Earnings	Employed	College
	(1)	(2)	(3)	(4)	(5)	(6)
High autonomy sample	.069***	.105***	.119***	.048*	-.021	.052***
	(.026)	(.027)	(.033)	(.028)	(.013)	(.015)
Low autonomy sample	.002	-.033	.041	-.063	-.005	-.044
	(.035)	(.039)	(.040)	(.046)	(.009)	(.036)

Notes: Table replicates the baseline estimation for adult skills, labor market and education outcomes on two split samples for high and low school autonomy. School autonomy information is taken from Hanushek et al. (2013). Reform countries with a high school autonomy are Poland, Czech Republic, Japan, Korea, Netherlands, Sweden, and United States while the low school autonomy reform countries are Canada, Germany, Greece, and Slovak Republic. From countries which did not change central exam status, Denmark, Finland, Israel, Ireland, Italy, New Zealand, United Kingdom are high autonomy countries and Turkey, Chile, Austria, Belgium, Norway, Spain are low autonomy countries. Robust standard errors clustered by graduation year × country in parentheses. * $p<0.10$, ** $p<0.05$, *** $p<0.01$

Table 12: Time to effect

Outcome	Numeracy skills		Literacy skills		Problem-solving skills	
	(1)	(2)	(3)	(4)	(5)	(6)
Central exams	.070***	.023	.070***	.014	.096***	.030
	(.021)	(.025)	(.022)	(.028)	(.022)	(.031)
Central exams 2+		.063*		.059*		.067*
		(.033)		(.032)		(.039)
Central exams 5+		-.006		.021		.030
		(.030)		(.029)		(.031)
Countries	27	27	27	27	23	23
Observations	131,398	131,398	131,398	131,398	88,827	88,827
R^2	.265	.265	.261	.261	.242	.242

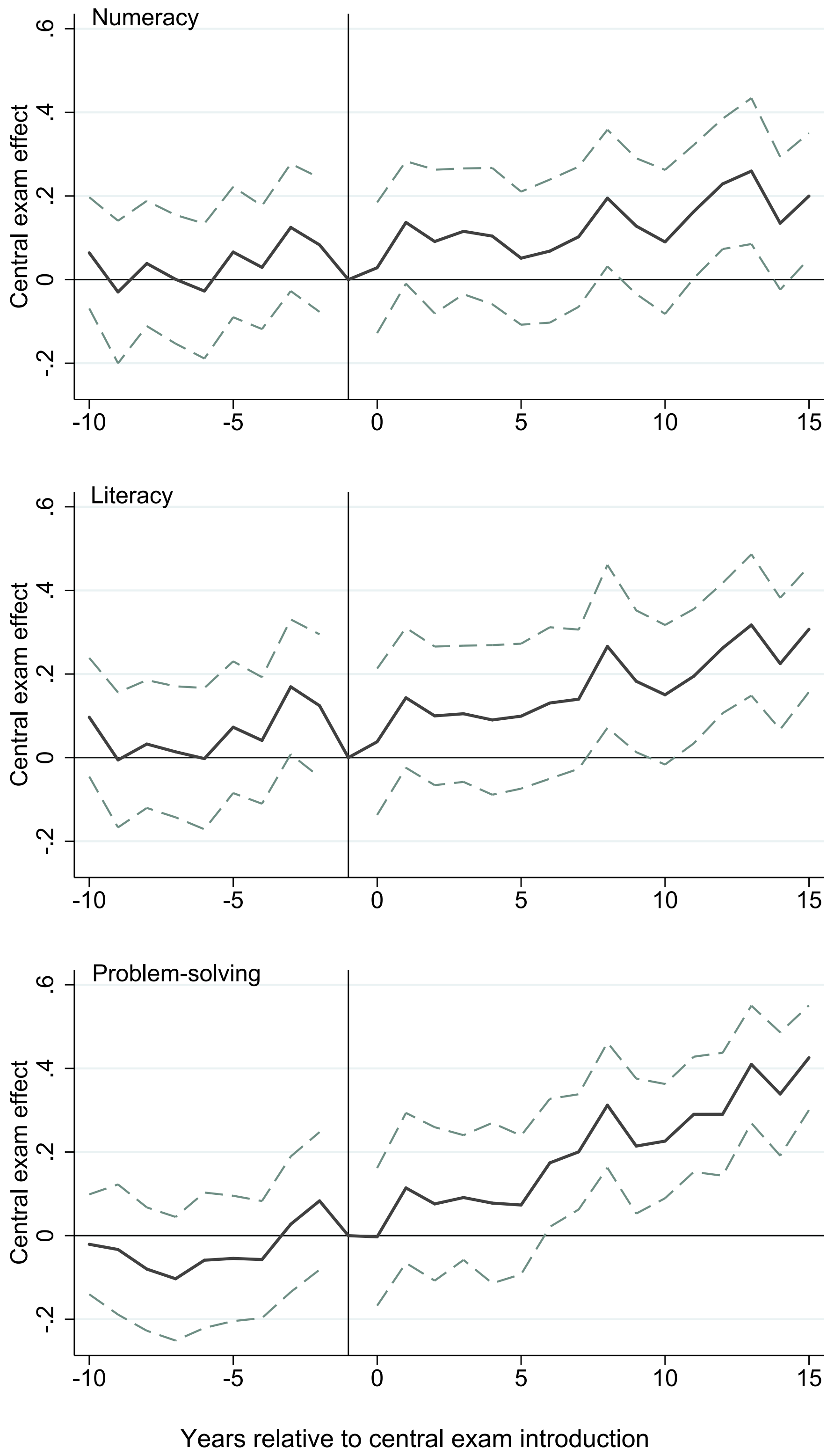
Notes: Estimates in columns (1), (3), and (5) correspond to our baseline specification, of Table 3, columns (4), (6), and (8), for a sample without the three countries with federal education systems, i.e. Canada, Germany, and the United States. In columns (2), (4), and (6) additional treatment indicators for "more than 2 years after the reform" (Central exams 2+) and for "more than 5 years after the reform" (Central exams 5+) are included. All estimates are weighted by adjusted sampling weights that give same weight to each country. Robust standard errors clustered by graduation year × country in parentheses. * p<0.10, ** p<0.05, *** p<0.01

Table 13: Comparison with correlations in PISA and TIMSS

	PIAAC (1)	PISA (2)	PIAAC (3)	TIMSS (4)
Central exams	.074	.175	.262	.361
	(.140)	(.124)	(.244)	(.290)
Countries	18	18	13	13
Observations	1,899	46,856	1,067	51,957
R^2	.018	.013	.033	.036

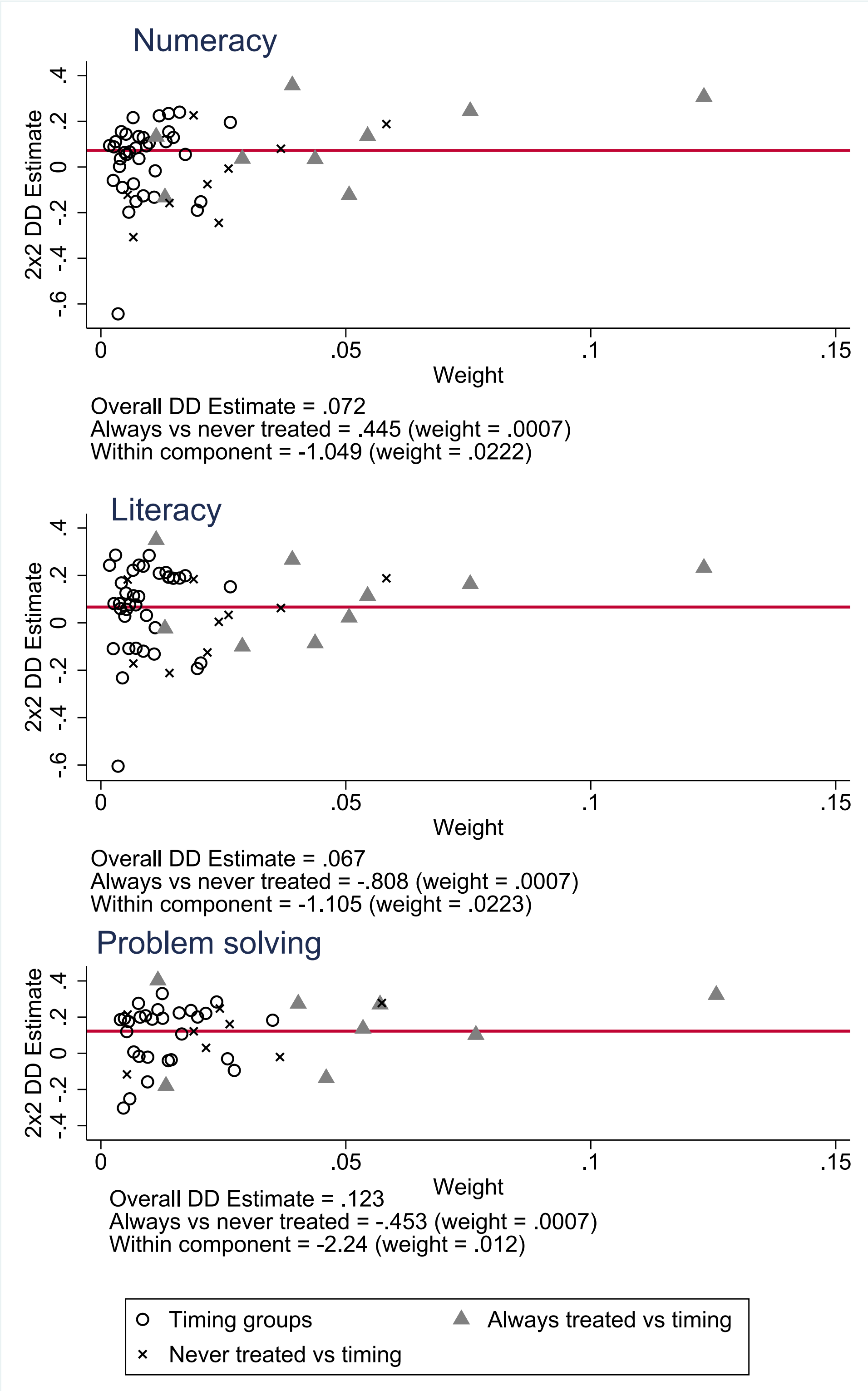
Notes: Dependent variable: globally standardized numeracy score for PIAAC and standardized math scores for PISA and TIMSS. PISA data refers to the PISA 2000 wave. TIMSS data refers to the 1999 TIMSS-Repeat study. The estimations in the PIAAC columns are based on a sample of PIAAC individuals who were 15 years old in 2000. The estimations in all columns except (1) are limited to the selection of countries which are common to PIAAC/PISA and PIAAC/TIMSS. All estimations are weighted by adjusted weights, where each country's weights sum up to 1. Robust standard errors (in parentheses) clustered by country. * p<0.10, ** p<0.05, *** p<0.01

Figure A1: Event study graph of the central exam effects on adult skills



Note: Depicted are 10 leads and 15 lags of the leads-and-lags equation: $y_{icg} = \sum_{\tau} \beta_{\tau} CE^{\tau}_{cg} + X'_{icg}\gamma + \delta_c + \theta_g + \varepsilon_{icg}$ along with their 95% confidence intervals. CE^{τ}_{cg} is a dummy equal to 1 if graduation year t is τ years after country c introduced central exams. Sample of 16 countries: 5 never-treated and 11 reform countries (see Table 1). The STATA command `eventdd` by Clarke and Tapia-Schythe (2021) is used to plot the event graph and estimate the underlying leads and lags. The first lead (-10) and the last lag (15) are accumulated, while the lead (-1) is normalized to zero.

Figure A2: Goodman-Bacon decomposition of central exam effects on adult skills



Note: Figure plots 2×2 DD components from the Goodman-Bacon decomposition against their weight for analysis of central exam effects on each domain of adult skills. The open circles are the timing-only terms. The closed triangles are terms in which one timing group acts as the treatment group and central exam countries act as the control group. The x's are terms in which one timing group acts as the treatment group and non-central exam countries act as the control group. The figure notes the average DD estimate and total weight on each type of comparison.

Table A1: Introduction of central exit examinations (CEE) at the upper secondary level (ISCED 3) across PIAAC countries

Country	Year of CEE-Introduction	Source
Austria	2016	Ministerium für Bildung und Frauen Österreich (BMBF): https://www.bmbwf.gv.at/Themen/schule/schulpraxis/zentralmatura.html, Hörner et al. (2007)
Belgium	-	Hörner et al. (2007), Eurydice Country Report: https://eacea.ec.europa.eu/national-policies/eurydice/content/assessment-second-and-third-stage-secondary-education-0_en
Canada*	1929-2000	General Accounting Office, Washington, 1993: http://files.eric.ed.gov/fulltext/ED361377.pdf, https://archive.org/details/ERIC_ED361377, Klinger (2008), Volante (2006), Bishop (1999a), WENR Country Report: https://wenr.wes.org/2017/09/education-in-canada
Chile	-	OECD (2010), OECD Education Policy Outlook Snapshot: https://www.oecd.org/education/highlightsChile.htm, WENR Country Report: https://wenr.wes.org/2013/12/introduction-to-the-higher-education-system-of-chile
Cyprus	2006	Hörner et al. (2007), Lamprianou (2012), Ministry of Education: http://www.moec.gov.cy/ypexams/en/examinations.html#:~:text=Pancyprian%20Examinations%20The%20objective%20of%20the%20Pancyprian%20Examinations,Public%20Higher%20Education%20Institutions%20of%20Cyprus%20and%20Greece, Eurydice Country Report: https://eacea.ec.europa.eu/national-policies/eurydice/content/assessment-general-upper-secondary-education-11_en
Czech Republic	2011	Greger and Kifer (2012), Eurydice Country Report: https://eacea.ec.europa.eu/national-policies/eurydice/content/assessment-upper-secondary-education-4_en
Denmark	1908	Bishop (1999a), Hörner et al. (2007), Eurydice Country Report: https://eacea.ec.europa.eu/national-policies/eurydice/content/assessment-general-upper-secondary-education-12_en

Notes: *In Canada the central exam regime differs across provinces. See Table A2 for information on central exam reform years across them. Online sources last retrieved August 18th, 2021.

Table A1: *Continued*

Country	Year of CEE-Introduction	Source
Estonia	1997	Archimedes Report 2010: https://eca.state.gov/files/bureau/estonia.pdf, Ministry of Education: https://www.hm.ee/en/activities/external-evaluation/state-examinations Education Youth Board Estonia: https://www.educationestonia.org/about-education-system/ Eurydice Country Report: https://eacea.ec.europa.eu/national-policies/eurydice/content/assessment-general-upper-secondary-education-14_en
Finland	1921	Matriculation Examination Board Finland: https://www.ylioppilastutkinto.fi/en/matriculation-examination/history Eurydice Country Report: https://eacea.ec.europa.eu/national-policies/eurydice/content/assessment-general-upper-secondary-education-15_en
France	1808	Hörner et al. (2007), Ministry of Education France: http://www.education.gouv.fr/cid60987/bac-2015-questions-reponses.html Eurydice Country Report: https://eacea.ec.europa.eu/national-policies/eurydice/content/assessment-general-upper-secondary-education-17_en
Germany*	1945-2008	Klein et al. (2009), Lüdemann (2011), Eurydice Country Report: https://eacea.ec.europa.eu/national-policies/eurydice/content/assessment-general-upper-secondary-education-21_en
Greece	1983, abolished in 2005	Polydorides (1986), Goulas and Megalokonomou (2021), Eurydice Country Report: https://eacea.ec.europa.eu/national-policies/eurydice/content/assessment-general-upper-secondary-education-23_en

Notes: * In Germany the central exam regime differs across federal states. See Table A2 for information on central exam reform years across them. Online sources last retrieved August 18th, 2021.

Table A1: *Continued*

Country	Year of CEE-Introduction	Source
Ireland	1927	State examination Commission Ireland: https://www.examinations.ie/?l=en&mc=li&sc=li Department of Education and Skills: http://www.education.ie/en/The-Education-System/Post-Primary/
Israel	1948	Syrquin (1997), Lavy (2008), Yemini et al. (2014)
Italy	1923	Nardi (2001), Eurydice Country Report: https://eacea.ec.europa.eu/national-policies/eurydice/content/assessment-general-upper-secondary-education-26_en
Japan	1979	Eckstein and Noah (1989), Watanabe (2013) Entrich (2015), Metzger et al. (2010), Ministry of Education: https://www.mext.go.jp/component/english/__icsFiles/afieldfile/2011/03/07/1303013_005.pdf WENR Country Report: https://wenr.wes.org/2021/02/education-in-japan
Korea	1994	Chang (2009), Korea Institute of Curriculum and Evaluation: http://www.kice.re.kr/sub/info.do?m=0205&s=english, WENR Country Report: https://wenr.wes.org/2018/10/education-in-south-korea, Jones (2013)
Lithuania	2002	Bethell and Zabulionis (2000), OECD (2002), Eurydice Country Report: https://eacea.ec.europa.eu/national-policies/eurydice/content/assessment-general-upper-secondary-education-31_en

Notes: Japan and Korea have national university entrance exams that we consider to be equivalent to a central exam in our main specification. Israel introduced "Bagrut" exams already in 1923 during the British Mandate of Palestine (1920-1948). Online sources last retrieved August 18th, 2021.

Table A1: *Continued*

Country	Year of CEE-Introduction	Source
Netherlands	1968	Hörner et al. (2007), Boezerooy and Huisman (2000), Government of the Netherlands: https://www.government.nl/topics/examination-in-secondary-education/secondary-school-leaving-examination, Eurydice Country Report: https://eacea.ec.europa.eu/national-policies/eurydice/content/assessment-general-lower-secondary-education-25_en, WENR Country Report: https://wenr.wes.org/2018/12/education-in-the-netherlands
New Zealand	before 1960	Hall (2000), Strachan (2002), New Zealands Qualifications Authority: https://www.nzqa.govt.nz/ncea/ncea-exams-and-portfolios/external/, www.nzqa.govt.nz/qualifications-standards
Norway	before 1960	Hörner et al. (2007), Organization for Internationalization in Education: https://www.epnuffic.nl/en/publications/find-a-publication/education-system-norway.pdf, Eurydice Country Report: https://eacea.ec.europa.eu/national-policies/eurydice/content/assessment-general-upper-secondary-education-39_en
Poland	2005	The Central Examination Commission: https://cke.gov.pl/en/egzamin-maturalny/, http://www.cke.gov.pl/images/stories/English/the_matura_exam.pdf
Singapore	1975	Statement Singapore Examinations and Assessment Board: https://www.seab.gov.sg/home/examinations/gce-a-level/about-gce-a-level Bishop (1997), Lim and Tan (1999)

Notes: Online sources last retrieved August 18th, 2021.

Table A1: *Continued*

Country	Year of CEE-Introduction	Source
Slovak Republic	2005	National Institute for Certified Educational Measurement: https://www.nucem.sk/en/measurements/maturita https://www.nucem.sk/en/measurements/maturita/brief-history-of-maturita Eurydice Country Report: https://eacea.ec.europa.eu/national-policies/eurydice/content/assessment-upper-secondary-education-11_en
Slovenia	1994	The National Examinations Centre: https://www.ric.si/general_matura/general_information/, Ilc et al. (2014) Eurydice Country Report: https://eacea.ec.europa.eu/national-policies/eurydice/content/assessment-general-upper-secondary-education-56_en
Spain	-	Bishop (1999a), Fuchs and Woessmann (2007), Hörner et al. (2007)
Sweden	abolished in 1968	Bishop (1999a), Hörner et al. (2007) Eurydice Country Report: https://eacea.ec.europa.eu/national-policies/eurydice/content/assessment-upper-general-and-vocational-secondary-education_en
Turkey	-	Yildirim et al. (2007), Measuring, Selection and Placement Center Turkey: http://www.osym.gov.tr/belge/1-2706/tarihsel-gelisme.html
United Kingdom	1951	WENR Country Profile: https://wenr.wes.org/2014/02/a-guide-to-the-gce-a-level http://www.a-levels.co.uk/history-of-a-levels.html A-Levels.co.com: http://www.a-levels.co.uk/history-of-a-levels.html
United States*	2003	Caves and Balestra (2018), Education Commission of the States: https://www.ecs.org/high-school-graduation-requirements/, https://www.ecs.org/state-education-policy-tracking/ WENR Country Profile: https://wenr.wes.org/2018/06/education-in-the-united-states-of-america

Notes: *In the United States the education system is decentralized. According to the definition of Bishop (1997), only the state of New York has central exams since 2003 in place. See also in Table A2. Online sources last retrieved August 18th, 2021.

Table A2: Central exams status in reform countries with federal education systems 1960-2015

Canada:				
Alberta (1984)	British Columbia [1974] (1983)	Manitoba [1970] (1991)	Newfoundland/ Labrador (1974)	New Brunswick (1994)
Northwest Territories[a] (2000)	Nova Scotia (1972)	Nunavut[a] (2000)	Prince Edward Island [1970]	Yukon[a] (2000)
Germany:				
Berlin (2006)	Brandenburg (2005)	Bremen (2007)	Hamburg (2005)	Hesse (2007)
Mecklenburg-Western Pommerania (1991)	Lower Saxony (2006)	North-Rhine Westphalia (2007)	Saxony (1993)	Saxony-Anhalt (1993)
Schleswig-Holstein (2008)	Thuringia (1990)			
United States:				
New York (2003)				

Notes: Years in round/square brackets indicate the year in which central exams were introduced/abolished. [a]Northwest Territories, Nunavut, and Yukon adopted the CEE system of British Columbia and Alberta around the year 2000. See Table A1 for a detailed list of sources.

Table A3: Coding of central exams in Fuchs and Woessmann (2007) compared to our coding across PIAAC/PISA and PIAAC/TIMMS countries

Country	F&W	Our	Dataset
Austria	0	0	PISA
Belgium	0	0	PISA/TIMSS
Canada	0.5	0.537	PISA/TIMSS
Chile	-	0	TIMSS
Cyprus	-	0	TIMSS
Czech Republic	1	0	PISA/TIMSS
Denmark	1	1	PISA
Finland	1	1	PISA/TIMSS
France	1	1	PISA
Germany	0.4	0.424	PISA
Greece	0	1	PISA
Ireland	1	1	PISA
Israel	-	1	TIMMS
Italy	1	1	PISA/TIMMS
Japan	1	1	PISA/TIMSS
Korea	1	1	PISA/TIMSS
Lithuania	-	0	TIMSS
Netherlands	1	1	PISA/TIMSS
New Zealand	1	1	PISA/TIMSS
Norway	1	1	PISA
Poland	1	0	PISA
Singapore	-	1	TIMSS
Slovak Republic	-	0	TIMSS
Slovenia	-	1	TIMSS
Spain	0	0	PISA
Sweden	0.5	0	PISA
Turkey	-	0	TIMSS
United Kingdom	1	1	PISA/TIMSS
United States	0.1	0.067	PISA/TIMMS