

Do High-Stakes Exams Promote Consistent Educational Standards?

Jack Rossiter, Might K. Abreh, Aisha Ali, and Justin Sandefur

Abstract

Each year over two million secondary-school students across English-speaking West Africa sit coordinated exams, with the explicit goal of maintaining consistent educational standards across schools and over time. Nevertheless, pass rates fluctuate from year to year, fueling speculation about cheating and short-term effects of education policies. To test these hypotheses, we construct an item bank of past exam questions spanning 2011-2019, and administer a hybrid test to 4,380 Ghanaian students. Scores across math items drawn from different exam years—when taken by an identical group of students on the same day—closely track fluctuations in Ghana’s national pass rates over time, absent any role for cheating or changes in real performance. Large swings in exam difficulty have significant implications for fairness and efficiency: half of candidates who failed to pass the maths test in 2015 would have passed in 2019.

View the online appendix for this paper: <https://www.cgdev.org/sites/default/files/do-high-stakes-exams-promote-consistent-educational-standards-online-appendix.pdf>

Keywords: High-stakes exams, student performance, secondary education, West Africa

JEL: I21, I25, I28, J24, O15, O55



Do High-Stakes Exams Promote Consistent Educational Standards?

Jack Rossiter
Center for Global Development
Corresponding author: jrossiter@cgdev.org

Might K. Abreh
Institute for Educational Planning and Administration, University of Cape Coast, Ghana

Aisha Ali
Center for Global Development

Justin Sandefur
Center for Global Development

Declarations of interest: none. Our particular thanks to Francis Amedahe for his substantial input to the research design and to seminar participants from the West African Exams Council. We thank Abdel Karim Fuseini, Allan Philip Barku, Baaba Sampson, Clemence Ayekple, Cyprian Ekow, Francis Ansah, James Amoateng, and Rita Denning who provided excellent research support. We also thank Abhijeet Singh, Alexis Le Nestour, Barbara Bruns, Caine Rolleston, Newman Burdett, and William Smith for their helpful comments. The views expressed here should not be attributed to the Center for Global Development or its funders. All remaining errors are our own.

The Center for Global Development is grateful for contributions from the Bill & Melinda Gates Foundation in support of this work.

Jack Rossiter, Might K. Abreh, Aisha Ali, and Justin Sandefur, 2021. "Do High-Stakes Exams Promote Consistent Educational Standards?" CGD Working Paper 581. Washington, DC: Center for Global Development. <https://www.cgdev.org/publication/do-high-stakes-exams-promote-consistent-educational-standards>

The data used in this paper is available here: <https://www.cgdev.org/sites/default/files/rossiter-et-al-WASSCE-exam-data-code.zip>. More information on CGD's research data and code disclosure policy can be found here: www.cgdev.org/page/research-data-and-code-disclosure.

Center for Global Development
2055 L Street NW
Washington, DC 20036

202.416.4000
(f) 202.416.4050

www.cgdev.org

The Center for Global Development works to reduce global poverty and improve lives through innovative economic research that drives better policy and practice by the world's top decision makers. Use and dissemination of this Working Paper is encouraged; however, reproduced copies may not be used for commercial purposes. Further usage is permitted under the terms of the Creative Commons License.

The views expressed in CGD Working Papers are those of the authors and should not be attributed to the board of directors, funders of the Center for Global Development, or the authors' respective organizations.

1 Introduction

Across Nigeria, Ghana, Sierra Leone, Liberia, and the Gambia, students seeking a secondary-school degree sit a coordinated test known as the West African Senior School Certificate Examination (WASSCE). Similar high-stakes exams are used in a majority of African nations, as well as many other countries around the world (Bashir et al. 2018). These exams serve an explicit dual mandate. On the one hand, stakes are high for students, as the exam determines university admissions and job placements and so, to some degree, must be unpredictable. At the same time, the exam is explicitly intended to maintain consistent standards for the educational *system*, and results are used by governments and civil society to monitor trends in educational performance. This requires comparability over time.

Theory suggests these two mandates are in tension (Neal 2013). Comparability generates predictability, gaming, and teaching to the test, all of which undermine the reliability of the exam for allocating university admissions and jobs. Despite this incompatibility, high-stakes “curriculum-based” examination systems, used for both screening and performance monitoring, have come to dominate educational assessment (Kolen and Brennan 2014). Proponents among policymakers and assessment experts argue that they help to maintain consistent educational standards (Dillard 2003) and secure educational outcomes in core subject areas (Phelps 2012). In policy dialogue they are presented as the “fairest and best system available to us” for filtering among candidates (U.K. House of Commons 2020), and for increasing opportunities among students from disadvantaged backgrounds (Kellaghan and Greaney 2019).

The ideals of fairness and consistent standards were cited as key motivations for the 1952 establishment of the West African Examinations Council (WAEC) which administers the WASSCE exam. The council retains the explicit goal, still advertised on its website, of ensuring that degrees in West Africa “[do] not represent lower standards of attainment than equivalent certificates of examining authorities in the United Kingdom.”¹ While WAEC purports to maintain fixed, absolute competency standards over time, in recent years WASSCE results jump from one year to the next. On the core mathematics test, which is the focus of our analysis in this paper, the pass rate in Ghana has swung from 82 percent in 2012, down to 54 percent in 2015, and back up to 86 percent in 2019 (Figure 2).

¹<http://www.waecgh.org/about-us>

To investigate the reasons for fluctuations in exam pass rates over time, we use methods from psychometrics to estimate exam difficulty over time. We administer hybrid test booklets containing questions from each WASSCE between 2011 and 2019 to a single sample of Ghanaian secondary school students. By concurrently grading responses from 4,380 Senior High School students we have a way of converting scores from one exam to those of any other exam in the period covered (Kolen and Brennan 2014; Patel and Sandefur 2020; IEA 2018). Applying methods from item response theory (IRT), we show how item differences have large impacts on the share of students reaching important grade boundary marks each year.

We find that most observed fluctuations are spurious, with differences in test difficulty explaining more than 80 percent of the variance in passes awarded. We present pass-rate estimates of between 48 percent and 85 percent in Maths, based on differences in examination difficulty alone. We use these data to look at the real trend in performance over time and estimate that around half of candidates who failed to pass the maths test in 2015 would have passed in 2019. Finally, we extend our empirical work with a content analysis of the 1,077 Maths and English items included in the study, to provide supportive evidence for variation in examination difficulty, based on the composition of items each year.

West Africa’s experience illustrates the potential pitfalls of pursuing a dual mandate in high-stakes examinations, with the imperative for a test that is unpredictable undermining the imperative for consistent standards. We conclude that large, year-to-year fluctuations in grade awards are not the consequence of real shocks (e.g., recessions or teacher strikes) or short-term effects of policy reforms, as they are commonly treated in public policy debates (Abreh, Owusu, and Amedahe 2018). Nor are they evidence of cheating by exam takers. Rather, the reason for wildly divergent results across years stems from inconsistencies in how test papers—and the standardisation processes that have grown around them—are designed and implemented.

Our results contribute to academic and policy debates on whether high-stakes exams are, in practice, beneficial for students (Smith 2016). Advocates of high-stakes testing point to its positive influence on incentives and the motivation of students (Gneezy et al. 2019), teachers (Glewwe, Ilias, and Michael Kremer 2010), parents (Bergbauer, Hanushek, and Woessmann 2018) and schools (Braun 2004). Critics of high-stakes testing highlight negative consequences including the enormous influence tests have on what is taught and how it is delivered (Jacob 2005) and the targeting of instruction to marginal students (Ballou and Springer 2017). If the

WASSCE does not reliably measure what pupils ought to be learning, it may reward the wrong things in classrooms (Woessmann 2018) and undermine educational standards.

This paper also raises questions for policymakers about how to measure progress in national education systems. As Ghana embarks on a major reform to Senior High School—“Free SHS”—political pressure for results is high (Graham, Van Gyampo, and Tuokuu 2020). While a substantial empirical literature, mostly from the United States, suggests that political pressures can quickly compromise assessment standards (Barlevy and Neal 2012; Campbell 1976), our findings suggest that it is inconsistent exam standards that have the potential to prompt spurious policy inference and inefficient government action (Singh 2020).

The remainder of the paper proceeds as follows. Section 2 describes our analytical approach. Section 3 describes the data collection for hybrid tests. Section 4 presents the main empirical results. Section 5 concludes.

2 Analytical approach

One approach to investigating the causes of the enormous variation in WASSCE pass rates shown in Figure 2 would be to contrast changes in school and student ‘inputs’ with changes in achievement. However, there are several variables that have not been measured consistently since 2011 and even where data have been obtained, research to date has struggled to explain observed swings in results using this approach (Abreh, Owusu, and Amedahe 2018).

An alternative approach is to start with data from previous examinations and see how performance on common items has varied. But, as with most public examinations, the WASSCE contains no overlapping items and is administered to non-equivalent groups of students each year. It is therefore impossible, with existing data, to determine whether performance varies in response to differences in the cohort or in the measurement instrument. To solve this, our analytical approach brings a *single* group of students to face items from several years, on the same day.

2.1 Converting Scores to a Common Scale

2.1.1 Estimating Average Examination Difficulty

We start from a random groups equating design, in which test-takers are randomly assigned the booklet to be administered (Kolen and Brennan 2014). When using this design, any differences between performance on a booklet is taken as a direct indication of difference in difficulty between those booklets, and multiple booklets may be equated at the same time. We extend this by creating common-item links across booklets, thereby combining features of the random groups and common-item non-equivalent group equating designs (Kolen and Brennan 2014). Items are randomly allocated to booklets, so that each ‘hybrid’ booklet contains items from every examination year and each item appears on at least two booklets. We end up with a web of item-booklet linkages, which we can later exploit to estimate item parameters. To complete the exercise, all booklet versions are used across all locations and a spiraling process is used to randomly allocate booklets to students.

Using this design we are able to calculate the difficulty of each item and, collectively, the average difficulty of each examination year. We estimate the following relationship:

$$Y_{ijkt} = \delta_j + \chi_k + \beta Year_t + \epsilon_{ijkt} \quad (1)$$

where Y_{ijkt} is the response to item i from candidate j on booklet k drawn from examination year t . Since individuals face items from multiple years, we include candidate fixed effects δ_j to control for differences in ability and since each booklet contains a unique set of items, we include booklet fixed effects χ_k to control for any influences from item location or item grouping across booklets. The average difficulty of each examination is obtained from coefficients on year fixed effects $Year_t$. The term ϵ_{ijkt} contains random error from each item response.

Equation (1) provides only a single parameter to capture the difficulty of each exam year. It makes no assumptions about the underlying ability distribution of students, and thus how pass rates will vary as exams get more or less difficult. A natural way to build in some assumptions about the underlying ability of students and thus the score distribution in any year, is to turn to Item Response Theory (IRT).

2.1.2 Estimating Pass Rates

Our equating design includes thousands of links between items, booklets and students which we use to concurrently calibrate parameters for every item included in the study (Wingersky and Lord 1984). Combining item parameters with an underlying ability distribution, we construct score distributions for each examination year and estimate the numbers of students that reach important grade boundaries. By setting fixed cutoff points, we look at difficulty-induced variation in pass rates over time and how this relates to grades awarded in historical data.

The main concept in IRT is the Item Characteristic Curve (ICC) which relates, for each item, a person's ability to the probability that they endorse the correct answer (Kolen and Brennan 2014). Among unidimensional models, the Three-Parameter logistic model (Birnbaum 1968) is the most general of the forms in widespread use.

$$P(y_i = 1|\theta_j) = c_i + (1 - c_i) \frac{\exp [Da_i(\theta_j - b_i)]}{1 + \exp [Da_i(\theta_j - b_i)]} \quad (2)$$

In this model, the functional form for an ICC is characterized by three item parameters, a_i , b_i and c_i , capturing item discrimination, item difficulty and a guessing parameter, respectively. The probability of endorsing the correct answer for that item i depends only on the student's ability θ_j and this set of parameters. A scaling constant D puts the trait scale in the same metric as the normal ogive model ($D = 1.7$) or in the metric of the logistic model ($D = 1$). A two-parameter model—which we use for all dichotomous items—can be derived from equation (2) by setting $c_i = 0$. Implicit in all models is a monotonicity assumption that as ability increases, the probability of endorsing a given item increases as well. As such, the higher the individual's ability, the higher is the probability of a correct response.

Our booklets are mixed-format, containing both dichotomously scored multiple choice items and polytomously scored constructed response items. The Generalised Partial Credit Model (GPCM) is used for all items that are scored in more than two ordered categories. An item scored $0, 1, \dots, m$ is divided into x adjacent logits and a positive response in category x implies a positive response to the preceding categories (Muraki 1992). The GPCM for an examinee with ability θ_j states that the probability of getting a score x on item i denoted by $P(y_i = x|\theta_j)$ is

$$P(y_i = x|\theta_j) = \frac{\exp \left[\sum_{v=1}^x Da_i(\theta_j - b_{iv}) \right]}{1 + \sum_{c=1}^m \exp \left[\sum_{v=1}^c Da_i(\theta_j - b_{iv}) \right]} \quad (3)$$

where D is a scaling constant as in equation (2), item i is described by a discrimination parameter a_i , and b_{iv} are m threshold parameters that represent the difficulty that distinguishes outcome v from the other outcomes in item i . In our application, we estimate up to seven threshold parameters for each constructed-response item. In the WASSCE, marks are awarded for constructed-response items by raters on a scale from 0 to 12. We are working with 117 such items from nine examination years, so to allow model convergence we reduce the number of threshold parameters to seven, at 1 and 2, 4, 6, 8, 10 and 12 marks. Operationally, this requires us to convert odd marks ≥ 3 to their nearest (higher) even number.² The selected IRT models are fit to the data via maximum likelihood estimation, returning all item and threshold parameters.

2.1.3 Assumptions

IRT models gain their flexibility by making well-known statistical assumptions, which may not hold in real testing situations (Kolen and Brennan 2014). In most IRT applications, items are trialed to assess function and model fit. In our administration we are restricted to using items that were developed by WAEC, which have already been fielded. Annex 2 provides evidence from an item fit analysis that assesses the adequacy of the functional form of the ICC implied by the chosen model (AERA and NCME 2014).

In the models used, each item is scored in two or more ordered categories and it is assumed that examinee ability is described by a single latent variable, θ , defined so that $-\infty < \theta < \infty$. The use of a single latent variable implies that the test construct is unidimensional. Research suggests that IRT equating is fairly robust to violations of the unidimensionality assumption which may arise when equating alternate forms of a test, so long as the violation is not too severe (Bolt 1999; Kolen and Brennan 2014). In support of the unidimensionality assumption, in this study we retain only items that were originally administered on paper and fielded to measure a single construct, either Mathematics ability or English Language ability (see section 3.1).

²Across all items, this increases marks awarded by mean 7.8 percent, standard deviation 1.1.

In applying IRT models, an assumption of local independence is also made, which means that after taking into account examinee ability, item responses are statistically independent. WAEC standards require test papers to be constructed so that all items are independent and no item gives a clue to answering another (WAEC 2019). We maintain this in hybrid forms and score dichotomous items separately. Among polytomous items, sub-parts often build on a common stimulus, so we score these at the item level.

3 Test Development and Data Collection

3.1 Constructing an item bank from past papers

For the purposes of this study, we develop an item bank containing WASSCE items from Terminal Assessments fielded between 2011 and 2019 in English and Mathematics. We focus on exams since 2011 because the Senior High School curriculum was reformed in 2010 and remains unchanged, despite review in 2020 (MOE 2018). For each subject there are several parts to the Terminal Assessment. We include everything that can be administered on paper without prior preparation. We exclude, for example, English sections on poetry, drama and prose which require knowledge of specific texts, which our respondents will not possess.

Specifically, we retain every item from Mathematics (Core) exams SC4021 and SC4022 and every item from English Language (Core) exam SC3021, Part A. SC4021 and SC4022 have a common structure from 2011 to 2019. SC4021 contains 50 multiple-choice items, each worth 1 mark. SC4022 is in two parts, Part I consisting of five constructed-response items, each worth 8 marks, and Part II with eight constructed-response items, each worth 12 marks. Candidates answer all items in Part I and select five items to answer from Part II. SC3021 Part A, known as ‘Lexis and Structure’, contains 50 multiple-choice items, each worth 1 mark. Prior to 2014, Part A covered the same content but included 70 items.

Items were obtained from past papers, released by WAEC. We worked with WAEC’s test development and test administration divisions to obtain digital copies of relevant papers since 2011. We obtained hard-copies of more recent tests from selected Senior High Schools and from exam-preparation books. With these inputs we were able to reproduce, word for word, all original examination items. Where graphics were used, we retained the original image.

Our final item bank contains 510 English items and 567 Mathematics items. As a share of the total marks available in these subjects, this represents 16 percent of the Terminal Assessment content for English and 100 percent of Terminal Assessment content for Mathematics, affecting how we analyse and present results in Section 4.

The bank can be used to provide a first impression of exam comparability, which we look at here for Maths. WAEC’s exam specification groups Maths items into seven content domains and five cognitive domains. Using these categories, in Figure 1a we map Maths content for 2011 to 2019 with markers scaled according to the total number of marks available for items at that location. Items that test number, algebra, mensuration and plane geometry are consistently prioritised each year, as expected by the examination specification. There is more variation in other domains, notably fewer marks available from trigonometry and statistics items in some years.

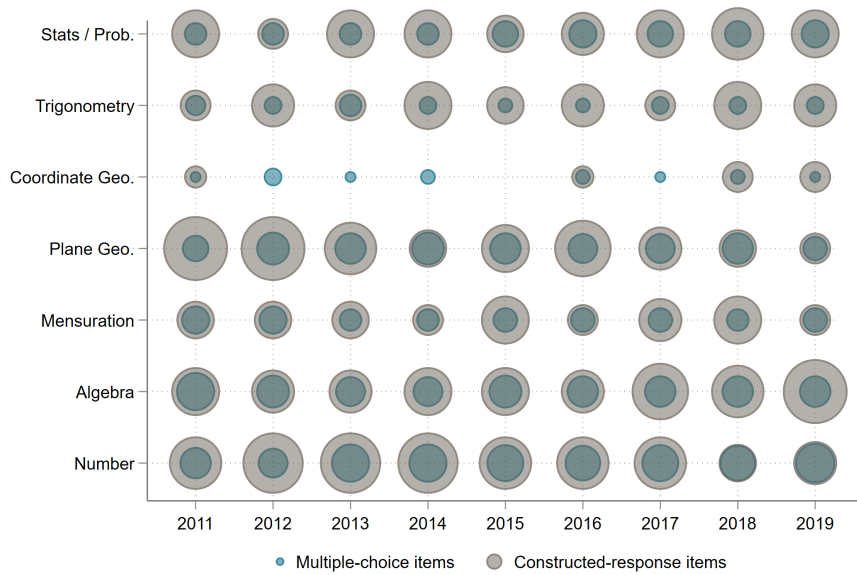
When items are grouped by cognitive domain, we see larger differences in exam construction that may influence difficulty, particularly for lower-ability students (Figure 1b). Each exam emphasises the ‘application’ of mathematical concepts to solve problems, mixed with a few items that test lower- and higher-order skills. There are, however, notable differences in cognitive domain coverage. For example, between 2014 and 2017 there were relatively few marks available at ‘Recall’ and ‘Comprehend’ levels, typically the levels at which less competent students can pick up marks.³ Instead, in these years and most prominent in 2017, there is a shift towards items requiring students to ‘Analyse’, which is, ex-ante, a more difficult cognitive domain.

3.2 Building Hybrid Tests

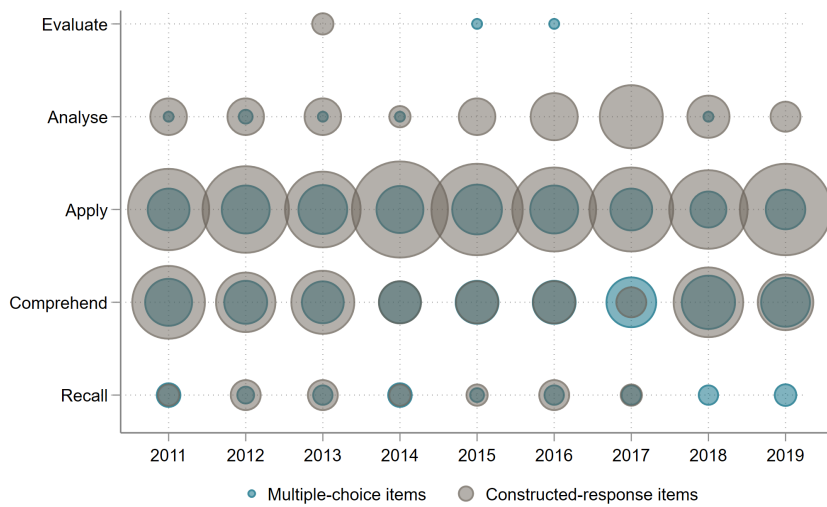
Hybrid test booklets are constructed with items randomly allocated to forms. We construct 18 unique English booklets, each containing 60 items, and 30 unique Mathematics booklets, each containing 38 items - 30 multiple choice, 8 constructed response. The number of items in each booklet is calibrated so that test length represents WASSCE standards.

Considerable effort was made to ensure that booklets represented original WASSCE papers.

³Although cognitive domains may increase in difficulty (i.e. it is easier to recall a concept than to apply that concept or to use multiple concepts to evaluate a scenario), this will not always correspond with empirical difficulty (recall of a rare term could easily be more difficult than the application of a basic law of maths).



(a) Marks categorised by year and content domain



Note: figures map items by year and content domain (above) and year and cognitive domain (below). At each point, two markers are shown, one for multiple-choice items, the other for constructed-response items. Markers are scaled according to the total number of marks available for items at that location.

(b) Marks categorised by year and cognitive domain

For both subjects, items are assigned within strata, so that each booklet retains the original content specification and item order is preserved. In addition, for English we stratify by content block, following those blocks used in original papers. Developing tests in this way results in forms that correspond to original tests in what they measure, with the only difference being the particular items that appear on the alternate forms.

3.3 Sample and Test Administration

Tests were administered to 4,380 students from 29 public Senior High Schools across 9 districts in Ghana’s Western, Central and Greater Accra regions. Students were all studying in Form 3, the final grade of Senior High School, and had a median age of 18. Over half of the students in the survey were female (see Annex A.1). On the direction of the Ghana Education Service, English and Mathematics exams were fielded at the same time to all students, on 18-19 December 2019.

Our analytical approach relies on our being able to distinguish between items in the bank, so that we can make claims about relative difficulty across years. It does not require a nationally representative sample of schools or students as we do not estimate population parameters. But it does require a sample of students that spans the full range of abilities, so that we avoid item-specific floor and ceiling effects. Avoiding floor effects is more difficult where learning levels are low (World Bank 2020). Not only that, but ours was a low stakes administration of an exam that exhibits high failure rates in most years. Our main concern was obtaining variation in performance on most-difficult items, which we achieved by oversampling historically higher-performing schools.

Public Senior High Schools in Ghana are organized into Category A, B, C and D in descending order of historical student performance and facility endowments, and this can be used as proxy for student ability (Ajayi 2015). Using a register of public schools obtained from the Ghana Education Service, we sampled schools within category, selecting 9 schools in Category A, 11 in Category B, and 9 in Category C. Schools were clustered in the nine districts to aid supervision during fieldwork: Abura/Asebu/Kwamankese, Agona West Municipal, Assin South, Cape Coast Metro, Mfantseman Municipal, Accra Metro, Tema Metro, Ellembele, Sekondi Takoradi Metro.

3.3.1 Marking Responses

Items fall into two categories: (i) objective-type items scored 0 or 1, for which responses are made on Optical Mark Recognition sheets and marked by machine with reference to the item-key; and (ii) subjective-type items, which are scored by markers, using WAEC mark rubrics for the relevant examination year.

Where students made no attempt at a multiple-choice item, their response is coded as missing. In some cases this is because the student did not reach that item. In other cases it is because the student was unable to answer the question and so they skipped it. In determining which items to count as incorrectly answered and which to count as not-administered we follow the approach taken by TIMSS (Foy and Yin 2015). Of the 450 objective-type items in the bank, 9 include an error in the stem which was introduced when constructing the item bank and hybrid tests. Marks for these items are omitted from the analysis.

Where students made no attempt at a constructed response item, markers left that as missing. Where students made an attempt at a constructed response item but failed to accrue any marks, that is scored 0. We follow the same TIMSS approach in coding constructed-response items that were not reached. The 117 constructed response items contain 468 parts: (a), (b), (c), (d). Ten sub-parts include an error introduced when constructing the item bank. These errors affect 47 marks out of 1,224 total, across all maths tests. All marks for affected items or sub-items have been omitted from the analysis (see Annex A.2).

4 Results

We present three main empirical findings. First, there are large and statistically significant differences in average WASSCE test difficulty. Second, this affects the proportion of students receiving ‘Pass’ (i.e. grades D7 and E8) or ‘Credit’ (grades A1 to C6) grades from year to year. Third, with exam difficulty estimates, we are able to explain a large share of the variation in grade awards over time. We begin with analysis for both subjects and then restrict our analysis to Maths when making comparisons with historical results.

4.1 Average Examination Difficulty

Across subjects and years, we identify large and statistically significant differences in average test difficulty. Table 4.1 shows coefficients on year dummies for 2011 through 2019. All results are estimated relative to 2015 for convenience, given that year's low performance in mathematics. Columns 1 and 4 show the additional marks per item for each examination year. Columns 2 and 5 include booklet fixed effects to account for possible influences from the grouping of items within booklets. Finally, our preferred results in columns 3 and 6 also include pupil fixed effects to control for any additional differences in the abilities of students that face each booklet.

In English, the average mark per item is 0.54 in 2015, falling between 0.49 (2019) and 0.56 (2016). Each item in this test is multiple choice so, on average, students provide correct responses for around half of the items. The minor changes from one year to the next correspond with the relatively stable trend we observe in WASSCE performance. WAEC changed the structure of English examinations in 2014 (Section 3.1), which aligns with the dip in pass grades shown in Figure 2. Converting estimated item difficulties into total marks would align with this, returning highest raw scores from 2011-2013.

In Maths, we see much larger differences in average item difficulty from year to year. In Column 6 we show an average mark per item of 0.65 in 2015, which falls to 0.62 marks in 2017 and rises to 0.97 marks in 2019. Average marks awarded are high in 2012, 2014 and 2019, corresponding with the pattern of performance in historical data. Based on these estimates, an average student that sits both the 2015 and 2019 exams would expect to see a difference of almost 50 percent in their raw score. There are two years, 2017 and 2018, which exhibit high performance in historical data but not in our analysis. Part of the explanation for the 2018 result is the omission of marks from high-value subjective-type items (Figure 4), which we address in the next section. It is less clear why average marks awarded for items from 2017 should be so low, but it is consistent with content coverage in that year (reviewed in Section 3.1).

Average difficulty is a useful indicator of differences across years, but provides a somewhat arbitrary point estimate for comparison. We are more interested in understanding how the combination of items affects performance at specific grade boundaries. We turn to this next.

Table 1: Average marks awarded per item, for each subject and examination year (our study)

	English			Mathematics		
	(1)	(2)	(3)	(4)	(5)	(6)
2011	-0.013*** (0.004)	-0.011*** (0.004)	-0.011*** (0.004)	0.095*** (0.017)	0.082*** (0.017)	0.082*** (0.017)
2012	-0.050*** (0.004)	-0.047*** (0.004)	-0.047*** (0.004)	0.137*** (0.017)	0.133*** (0.017)	0.137*** (0.016)
2013	-0.025*** (0.004)	-0.031*** (0.004)	-0.031*** (0.004)	0.078*** (0.017)	0.071*** (0.017)	0.070*** (0.016)
2014	-0.045*** (0.004)	-0.048*** (0.005)	-0.048*** (0.004)	0.120*** (0.017)	0.115*** (0.017)	0.112*** (0.016)
2016	0.022*** (0.004)	0.023*** (0.005)	0.023*** (0.004)	0.049*** (0.017)	0.037** (0.017)	0.036** (0.016)
2017	-0.033*** (0.004)	-0.028*** (0.005)	-0.028*** (0.004)	-0.017 (0.017)	-0.028 (0.017)	-0.027* (0.016)
2018	-0.049*** (0.004)	-0.045*** (0.005)	-0.045*** (0.004)	-0.010 (0.017)	-0.023 (0.017)	-0.026 (0.016)
2019	-0.044*** (0.004)	-0.050*** (0.005)	-0.050*** (0.004)	0.341*** (0.017)	0.326*** (0.017)	0.322*** (0.016)
Constant	0.541*** (0.003)	0.541*** (0.003)	0.541*** (0.003)	0.642*** (0.012)	0.651*** (0.012)	0.652*** (0.012)
Observations	257632	257632	257572	143171	143171	143171
Booklet FE	No	Yes	Yes	No	Yes	Yes
Pupil FE	No	No	Yes	No	No	Yes

Note: The primary unit of observation is a single test item response. Coefficients represent the differences in the average number of marks awarded per item, based on Equation (1), and relative to 2015 which we set as the base year. In English there were only multiple choice items whereas for Maths there are also constructed response items. For English, there were 70 total items each year until 2013, after which this fell to 50. For Mathematics, there were 63 total items in each year. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Standard errors are shown in parentheses.

4.2 ‘Pass’ and ‘Credit’ pass rates

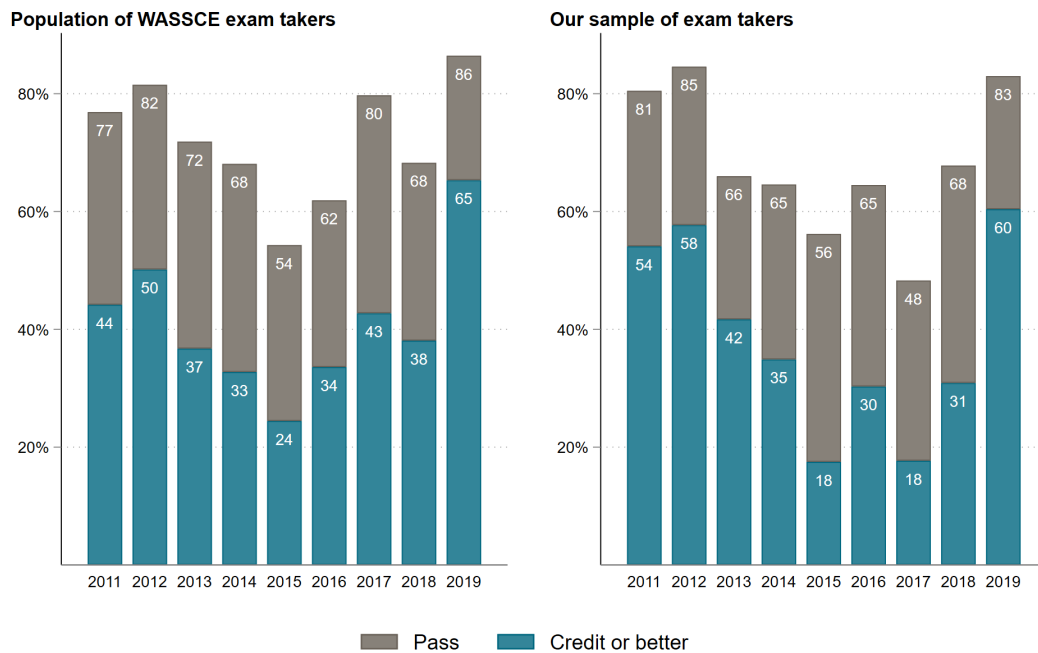
To evaluate the proportion of candidates reaching each grade boundary, a full distribution of scores is required. IRT’s item response function can be used to predict every candidate-item response from ability and item parameters. To do this, item discrimination and difficulty (for each response category) are estimated in a concurrent calibration using all survey data. Then a single group of 1,000 normally distributed candidate abilities is generated. For each candidate, the probability of correctly endorsing each response category is calculated, before total expected scores are calculated based on the probability of correct endorsement. Finally, the percentage correct is calculated for every examination year so that each candidate has nine scores, one for 2011 through 2019. In this conversion from raw scores to percentage correct we make the assumption that the 3 percent of marks that were previously omitted (Table 4) have the same average difficulty as other items in that examination year.

After WASSCE are marked, grade boundaries are set in a ‘Standard Fixing and Grade Award Meeting’. In this meeting, a panel of representatives from WAEC countries make their judgments using data on the current year’s mark distribution, copies of question papers and marking schemes, grade award details for the previous year’s examination, chief examiner’s reports, samples of student responses and several other inputs (WAEC 2019). In principle, the process of setting these grade boundaries is what allows examination councils to claim comparability over time (Newton 2007). We’re agnostic about the Standard Fixing process. Instead we use constant grade boundaries across examination years. By doing so we demonstrate that the Standard Fixing process appears to do very little to counter variation induced by changes in examination difficulty.

With grade boundaries fixed, we estimate the proportion of candidates that reach ‘Pass’ and ‘Credit’ levels each year. We set grade boundary marks by pooling ‘Pass’ and ‘Credit’ rates across years, representing a test of average difficulty. We find that differences in the composition of items in each exam have large impacts on each year’s score distribution. The share of students reaching fixed boundaries varies considerably, with pass-rates of between 48 percent and 85 percent. The main purpose of this exercise is to contrast difficulty-induced variation in performance with historical trends in grades awarded, which is now possible.

4.3 New evidence on past achievement trends

Figure 2 presents WAEC results alongside our difficulty-based estimates. Four features stand out. First, we track the early increase in ‘Pass’ and ‘Credit’ awards to 2012, the downward trend to 2015 and the subsequent rise to 2019. Second, we obtain very similar, although non-identical, rank order across examination years. Third, our data suggest that there should have been a larger share of ‘Credit’ awards, among all passes, in the years before 2015. Fourth, 2017 stands out as showing a large difference between WAEC results and our estimates.



Note: Credit or better = WASSCE grades A1-C6; Pass = WASSCE grades D7-E8

Figure 2: Actual WASSCE pass rates compared with difficulty-induced pass rates in our study

When adjusted for examination difficulty, and omitting 2017 from both WAEC results and study estimates, variance in ‘Pass’ awards falls by 88 percent and variance in ‘Credit’ awards falls by 71 percent. We split the historical variation in grade awards into two parts: one part that depends on test difficulty and another that represents cohort ‘quality’,⁴ which may be used

⁴There are several other potential sources of variation in grades awarded including changes in the contribution from Continuous Assessment, which we discuss in the next section.

to reconstruct a ‘true’ performance trend since 2011. In contrast with the dramatic fluctuations that the public and policymakers debate each year, we find a modest change in performance since 2011 (Figure 3a). At the ‘Pass’ level performance rises to 2013, then returns to roughly 2011 levels before an uptick in 2019. At the ‘Credit’ level, improvements are gradual and sustained through to 2019.

An implication of the dramatic changes in test difficulty is that the ability level required to pass the Maths WASSCE changes from year to year. We look at this with a thought experiment. Suppose that every candidate since 2011 had faced the same test, set at the average difficulty level of all tests in the period: how would the pass rate for the Maths WASSCE have changed in this scenario? To estimate this, we calculate what proportion of the change in performance is due to differences in test difficulty and look at what it would mean to eliminate this (Figure 3b).

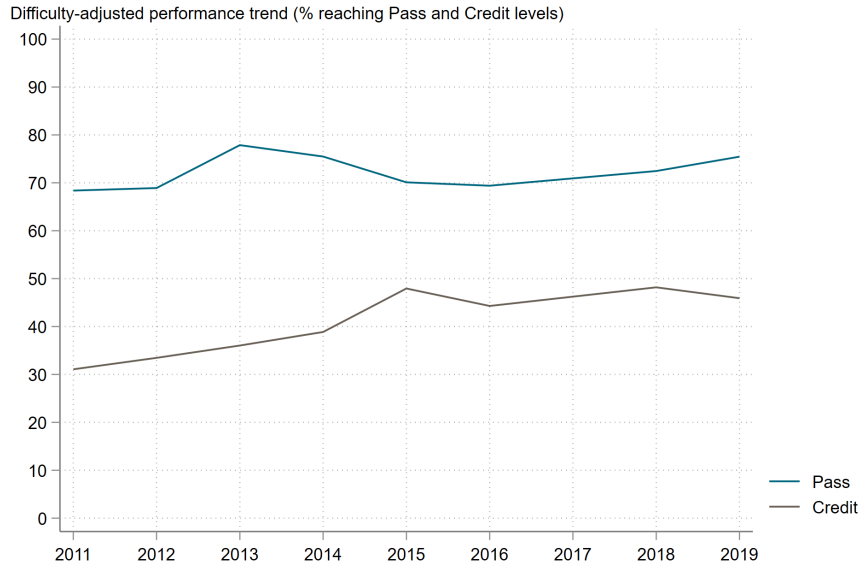
From this decomposition, we estimate the number of ‘excess’ failures and ‘excess’ passes in a given year (Table 2). In some years (2011, 2012, 2019) a particularly easy WASSCE benefits anywhere between 10,000 and 35,000 Ghanaian candidates. Conversely, between 2013 and 2018, the more difficult maths WASSCE led to excess failures of a similar magnitude or, equivalently, between 15 and 35 percent of all students that failed the maths WASSCE in any given year may have done so thanks to a particularly difficult test. When comparing across cohorts, around half of candidates who failed to pass the maths test in 2015 would have passed in 2019.

4.4 Limitations

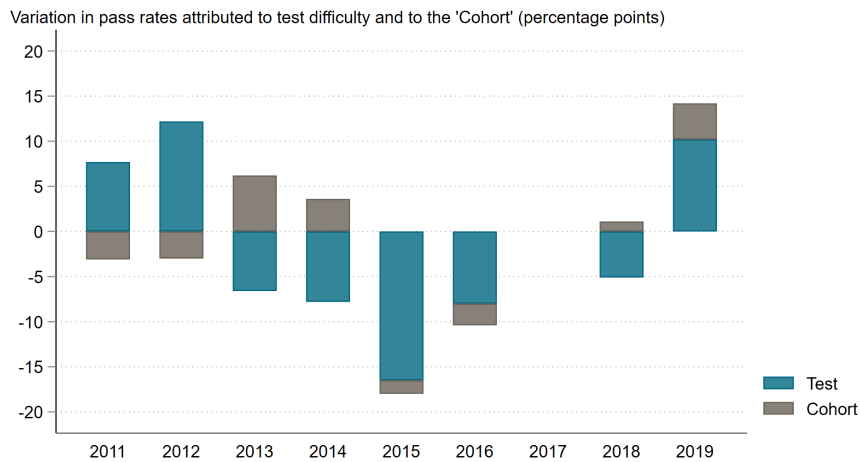
Our results cannot explain all of the changes in grades awarded over time. In addition to grade boundary adjustments (Section 4.2) there are at least four additional sources of variation.

(1) Students are awarded marks for school-based Continuous Assessment, which contribute to their WASSCE grade. We cannot reproduce continuous assessment scores, so our data cover 70 percent of the maths marks available each year. For English this falls to 11 percent of the marks available each year which is why we avoid strong claims about the relationship between overall grade awards and our difficulty estimates in that subject. Our analysis assumes that there is no systematic variation in Continuous Assessment scores from year to year.

(2) We do not capture any post-examination adjustments to the marking rubric for items



(a) Difficulty-adjusted trends in Mathematics performance



Note: we separate the differences in pass rates into two parts: one can be explained by test difficulty and one cannot. For convenience, we refer to this latter part as 'Cohort', indicating a genuine change in cohort performance but we cannot, with available data, test any hypothesis about cohort competency directly. All values are relative to an average difficulty test, based on pooled data for 2011-2019. We cannot, with our estimates of test difficulty, provide a reasonable explanation for the sharp increase in Pass grades awarded in 2017; so we omit that year from this figure.

(b) Decomposing variation in Mathematics Pass rates

Table 2: Difficulty-induced excess failures (and passes) in the maths WASSCE

	Exam candidates	Historical fail rate	Excess fails (% candidates)	Excess fails (candidates)
2011	147,092	23%	-7.7	-11,326
2012	172,913	18%	-12.2	-21,095
2013	406,001	28%	6.6	26,796
2014	237,683	32%	7.8	18,539
2015	262,913	46%	16.5	43,381
2016	268,187	38%	8.0	21,455
2017*	286,544	20%	24.6	
2018	312,486	32%	5.1	15,937
2019	342,529	14%	-10.2	-34,938
	2,436,348	28%		

Note: excess fails shows the proportion and number of additional students who we estimate would have failed (or passed) the Mathematics WASSCE each year, on the assumption that test difficulty had not varied over the period. All values are relative to a constant ‘average difficulty’ test, based on pooled data from 2011-2019. It represents the share of pass-rate variation that we can explain with changes in test difficulty (see Figure 3b). * We cannot explain the sharp increase in ‘Pass’ grades awarded in 2017 based on our information on test difficulty, so we omit any estimation of excess fails for this year. Source: historical sitters and grade awards from WAEC; excess failures based on authors’ calculations.

that examiners decided contained an error. Our analysis applies original WAEC mark rubrics and we are not privy to decisions made by the council which may have modified marks for individual items.

(3) Students may choose which items to answer in the second part of their Mathematics examination. Each candidate selects five questions from eight and more popular items will carry more weight in marks obtained. In our study, we fielded every item from each year and all responses from these contribute to the overall difficulty estimate.

(4) We cannot discount the possibility that the residual differences in grade awards (Figure 3a) result from genuine differences in student preparation. There have been several reforms to secondary education in Ghana over the past decade, which have sought to increase access and improve quality in low-performing schools.⁵ It is, however, beyond the scope of this paper to investigate or evaluate the impacts of Ghana's many education reforms on student outcomes.

By working with WAEC's historical data it would be possible to look at the importance of these factors. Without that, our results still demonstrate a tight relationship between exam difficulty and grade awards, suggesting that year-on-year variation in these factors plays only a minor role.

5 Conclusion

Many high stakes assessment systems are designed to produce consistently scaled scores that are comparable over time. Comparability comes at a cost. Tests use standardised question formats, fixed content coverage and, in many cases, common items to create a statistical link between papers. This makes each test predictable, and generates potentially undesirable incentives for students, teachers and officials (Jacob 2005; Ballou and Springer 2017).

Paradoxically, our results show that in Anglophone West Africa these negative consequences arise without the benefit of comparability. Pass rates fluctuate enormously from year to year. We show that differences in WASSCE exam difficulty explain more than 80 percent of the variance over time in maths grades awarded. As a result, education officials and civil society monitor progress using unreliable performance metrics.

⁵See, for example, the Ghana Secondary Education Improvement Project (World Bank Project P145741)

Seven in ten Ghanaian secondary school students expect to be a government employee or in a profession dominated by government employees by the age of 25 if they complete Senior High School (Duflo, Dupas, and Micheal Kremer 2019). An unreliable WASSCE exam has obvious implications both for fairness to candidates and for the efficient allocation of these jobs and of university admissions (Zimmerman 2014; Ozier 2018).

Our analysis is not without limitations. We cover a majority of information from terminal exams, but cannot account for continuous assessment and other information that WAEC uses to award grades. Even though we can't directly estimate the impact of each part of the process, our findings suggest that these are likely to have only a minor influence on grades awarded. We are also limited in our subject coverage, as our methodology is better suited to the objective content from the mathematics (and portions of the English) exam than to other subjects. But we have no reason to suspect comparability over time is higher on subjects with more subjective grading, and it may very well be lower.

We see no reason to suggest deliberate political interference with the exam. For one, the changes in performance from 2011-2019 appear fairly random and do not show steady inflation—as might be more common in assessment corruption (Barlevy and Neal 2012; Neal 2010)—nor any obvious relationship to political changes in the same period.

Several technical fixes involving more formulaic approaches to setting exam standards could, in theory, partially resolve the problems identified in this study: e.g., statistical alignment at grade boundaries, with expert judgment used only in a confirmatory capacity; the use of external reference tests to directly set standards; or the use of item response theory to make a direct link between tests (Burdett et al. 2013; Pointer 2014). Some of these technical fixes are less relevant, however, to other subjects and question types beyond mathematics. These technical fixes also fail to address the underlying challenge of using a single high-stakes assessment for both screening pupils and maintaining consistent educational standards. It may be preferable to develop separate assessment systems that are designed specifically for each measurement task Barlevy and Neal 2012.

Looking forwards, comparatively little attention has been paid to evaluating the integrity of public examinations, despite their widespread use and high stakes. Our approach could be quickly and relatively easily applied in different countries that follow the same assessment tradition.

References

- Abreh, Might Kojo, Kofi Acheaw Owusu, and Francis Kodzo Amedahe (2018). “Trends in Performance of WASSCE Candidates in the Science and Mathematics in Ghana: Perceived Contributing Factors and the Way Forward”. In: *Journal of Education* 198.1, pp. 113–123. ISSN: 25155741. DOI: 10.1177/0022057418800950.
- AERA, APA and NCME (2014). *Standards for Educational and Psychological Testing*. Washington D.C.: American Educational Research Association.
- Ajayi, Kehinde F (2015). “School Choice and Educational Mobility: Lessons from Secondary School Applications in Ghana”.
- Ballou, Dale and Matthew G Springer (2017). “Has NCLB encouraged educational triage? Accountability and the distribution of achievement gains”. In: *Education Finance and Policy* 12.1, pp. 77–106.
- Barlevy, Gadi and Derek Neal (2012). “Pay for percentile”. In: *American Economic Review* 102.5, pp. 1805–31.
- Bashir, Sajitha et al. (Sept. 2018). *Facing Forward: Schooling for Learning in Africa*. Vol. 91. The World Bank, pp. 399–404. ISBN: 978-1-4648-1260-6. DOI: 10.1596/978-1-4648-1260-6. URL: <http://elibrary.worldbank.org/doi/book/10.1596/978-1-4648-1260-6>.
- Bergbauer, Annika B, Eric A Hanushek, and Ludger Woessmann (2018). “Testing”.
- Birnbaum, Allan (1968). “Some Latent Trait Models and Their Use in Inferring an Examinee’s Ability”. In: *Statistical theories of mental test scores*. Ed. by F.M. Lord and M.R. Novick. Reading, MA: Addison-Wesley, pp. 397–479.
- Bolt, Daniel M (1999). “Evaluating the effects of multidimensionality on IRT true-score equating”. In: *Applied Measurement in education* 12.4, pp. 383–407.
- Braun, Henry (2004). “Reconsidering the Impact of High-stakes Testing”. In: *Education Policy Analysis Archives* 12.1.
- Burdett, Newman et al. (2013). *Maintaining qualification and assessment standards: summary of international practice*. Tech. rep. Slough: NFER.
- Campbell, Donald T (1976). “Assessing the impact of planned social change”. In: *Occasional Paper Series* 8.
- Dillard, Mary E (2003). “Examinations standards, educational assessments, and globalizing elites: the case of the West African Examinations Council”. In: *The Journal of African American History* 88.4, pp. 413–428.

- Duflo, Ester, Pascaline Dupas, and Micheal Kremer (2019). “The Impact of Free Secondary Education: Experimental Evidence from Ghana”. In: *Massachusetts Institute of Technology Working Paper Cambridge, MA*. 1254167, pp. 1–105.
- Foy, Pierre and Liqun Yin (2015). “Scaling the TIMSS 2015 achievement data”. In: *Methods and procedures in TIMSS*.
- Glewwe, Paul, Nauman Ilias, and Michael Kremer (2010). “Teacher incentives”. In: *American Economic Journal: Applied Economics* 2.3, pp. 205–27.
- Gneezy, Uri et al. (2019). “Measuring success in education: the role of effort on the test itself”. In: *American Economic Review: Insights* 1.3, pp. 291–308.
- Graham, Emmanuel, Ransford Edward Van Gyampo, and Francis Xavier Tuokuu (2020). “A Decade of Oil Discovery in Ghana: Implications for Politics and Democracy”. In: *Ghana Social Science Journal*.
- IEA (2018). *Rosetta Stone: measuring global progress towards SDG4 by linking assessments results to TIMSS and PIRLS International Benchmarks of Achievement*. International Association for the Evaluation of Educational Achievement.
- Jacob, Brian A (2005). “Accountability, incentives and behavior: The impact of high-stakes testing in the Chicago Public Schools”. In: *Journal of public Economics* 89.5-6, pp. 761–796.
- Kellaghan, Thomas and Vincent Greaney (2019). *Public Examinations Examined*. The World Bank.
- Kolen, Michael and Robert Brennan (2014). *Test Equating, Scaling, and Linking: Methods and Practices*. Third Edition. ISBN: 9780387712642. DOI: 10.1007/978-1-4939-0317-7.
- MOE (May 2018). *National Pre-tertiary Education Curriculum Framework*. Tech. rep. National Council for Curriculum and Assessment. Ministry of Education.
- Muraki, Eiji (1992). “A Generalized Partial Credit Model: Application of an EM Algorithm”. In: *Applied Psychological Measurement* 16.2, pp. 159–176. ISSN: 15523497. DOI: 10.1177/014662169201600206.
- Neal, Derek (2010). “Aiming for efficiency rather than proficiency”. In: *Journal of Economic Perspectives* 24.3, pp. 119–32.
- (2013). “The consequences of using one assessment system to pursue two objectives”. In: *The Journal of Economic Education* 44.4, pp. 339–352.
- Newton, Paul E. (2007). “Contextualising the Comparability of Examination Standards”. In: *Techniques for monitoring the comparability of examinations standards*. Ed. by P. Newton et al. London: Qualifications and Curriculum Authority, pp. 9–42.

- Ozier, Owen (2018). “The Impact of Secondary Schooling in Kenya: A Regression Discontinuity Analysis”. In: *Journal of Human Resources* 53.1, pp. 157–188. ISSN: 0022-166X. DOI: 10.3368/jhr.53.1.0915-7407R. URL: <http://jhr.uwpress.org/lookup/doi/10.3368/jhr.53.1.0915-7407R>.
- Patel, Dev and Justin Sandefur (2020). “A Rosetta Stone for Human Capital”. In: *Center for Global Development Working Paper* 550.
- Phelps, Richard P. (2012). “The Effect of Testing on Student Achievement, 1910–2010”. In: *International Journal of Testing* 12.1, pp. 21–43. DOI: 10.1080/15305058.2011.602920. URL: <https://doi.org/10.1080/15305058.2011.602920>.
- Pointer, William (2014). *Setting the grade standards in the first year of the new GCSEs*. Tech. rep. Manchester: QA Centre for Education Research and Practice.
- Singh, Abhijeet (July 2020). “Myths of Official Measurement : Auditing and Improving Administrative Data in Developing Countries”. In: *RISE Working Papers* 20/042.
- Smith, William C (2016). “The Global Testing Culture: shaping education policy, perceptions, and practice”. In: Symposium Books Ltd.
- U.K. House of Commons (Sept. 2020). “Oral evidence taken before the Education Committee on 16 September 2020, on Accountability hearings, HC 262”. In:
- WAEC (2019). *Manual of Procedures of Activities for Test Development Division*. Accra, Ghana: West African Examination Council.
- Wingersky, Marilyn S and Frederic M Lord (1984). “An investigation of methods for reducing sampling error in certain IRT procedures”. In: *Applied Psychological Measurement* 8.3, pp. 347–364.
- Woessmann, Ludger (2018). “Central exit exams improve student outcomes”. In: *IZA World of Labor* January, pp. 1–10. DOI: 10.15185/izawol.419.
- World Bank (2020). *The Human Capital Index 2020 Update: Human Capital in the Time of COVID-19*. Washington, D.C.: The World Bank.
- Zimmerman, Seth D. (2014). “The returns to college admission for academically marginal students”. In: *Journal of Labor Economics* 32.4, pp. 711–754. ISSN: 0734306X. DOI: 10.1086/676661.