The High Stakes of Bad Exams

Jack Rossiter, Might Kojo Abreh, Aisha Ali, Justin Sandefur *

Abstract

Each year two million secondary-school students across West Africa sit coordinated exams. Pass rates fluctuate enormously, fueling speculation about cheating and short-term policy changes. To investigate these hypotheses, we construct hybrid exams containing items spanning 2011-2019 and administer these to 4,380 students. Exam difficulty alone explains 80 percent of pass rate fluctuations in Ghana, while additional factors remain influential in Nigeria and elsewhere. Half of the candidates who failed mathematics in 2015 would have passed in 2019. Model based estimates imply that improving exam comparability would increase the Mincerian return to skills among secondary school graduates by 6 percentage points.

JEL Classification: I25, I26, J24, O15, O55

* Jack Rossiter, Center for Global Development (jrossiter@cgdev.org). Might Kojo Abreh, Institute for Educational Planning and Administration, University of Cape Coast. Aisha Ali, Center for Global Development. Justin Sandefur, Center for Global Development. The data used in this article are available online: Rossiter et al. 2023. "Replication Data for: The High Stakes of Bad Exams". Harvard Dataverse. https://doi.org/10.7910/DVN/VM5BOQ. Our particular thanks to Francis Amedahe and seminar participants from the West African Exams Council. We thank Abdel Fuseini, Allan 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, William Smith, and anonymous referees for their helpful comments. The views expressed here should not be attributed to the Center for Global Development, the Institute for Educational Planning and Administration, or their funders. All errors are our own. This work was supported by the Bill & Melinda Gates Foundation, Seattle, WA [grant number OPP1198125]. Other support summing to $10,000 in the past three years: Rossiter, Asian Development Bank, Echidna Giving, UK Foreign, Commonwealth and Development Office; Abreh, none; Ali, none; Sandefur, Asian Development Bank, Centre for Effective Altruism, Echidna Giving, UK Foreign, Commonwealth and Development Office, Pousaz Philanthropies, World Bank Group. No party had the right to review the paper prior to its circulation. IRB approval: University of Cape Coast Institutional Review Board, UCCIRB/EXT/2019/37.

doi:10.3368/jhr.0621-11739R1
This open access article is distributed under the terms of the CC-BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0) and is freely available online at: http://jhr.uwpress.org
I. Introduction

Almost 90 percent of countries use high stakes exams at the end of secondary school to award credentials and filter students into universities (Furuta 2021; Bashir et al. 2018). But economists have mostly ignored their implications for education systems and labor markets. 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). These exams serve an explicit dual mandate. On the one hand, the 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 screening and performance monitoring 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.” 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 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. This variation is not unique to Ghana, with the proportion of pass grades fluctuating from 71 to 93 percent in Nigeria, from 13 to 40 percent in The Gambia, and from 14 to 74 percent in Sierra Leone, in the study period.

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 for school candidates, fielded between 2011 and 2019, to a single sample of Ghanaian secondary school students. Among WAEC members, Ghana is the ideal context in which to study these effects because of its size (more than twice the number of candidates from Liberia, Sierra Leone and The Gambia, combined), average performance level and relative stability throughout the period, with less concern around widespread cheating than in Nigeria and without the shocks experienced in Sierra Leone. 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, 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 in Ghana are spurious, with differences in test difficulty explaining more than 80 percent of the variance in passes awarded. We then look at the real trend in performance over time and estimate that around half of candidates who failed to pass the mathematics test in 2015 would have passed in 2019. We go on to illustrate what fluctuations in exam difficulty imply elsewhere, with changes in test difficulty helping to explain up to 62 percent of the variation in credit-pass grades awarded in Nigeria since 2014. We extend our empirical work with a content analysis of the 1,077 English and mathematics items included in the study, to provide supportive evidence for variation in examination difficulty, based on the composition of items each year.

Besides their implications for fairness or student incentives, large fluctuations in exam difficulty may undermine productivity through the misallocation of talent. We sketch a simple framework to think about what our results mean for worker productivity and combine parameters from our mathematics tests with new estimates of the return to observable skills in the Ghanaian labor market. Our calculations suggest that improvements in the comparability of the WASSCE exam could increase the Mincerian return to skills by about 6 percentage points among workers completing secondary school.

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 in Ghana are not a 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 standardization processes that have grown around them—are designed and
implemented. These inconsistencies help to explain variation in performance outside Ghana, but the records of other WASSCE members suggest roles for other factors as well. For instance, Nigeria’s rapid improvement in official WASSCE exam performance from 2011 to 2014 also appears to depend on some combination of big changes in student quality and/or widespread cheating. Elsewhere, especially in Sierra Leone, major shocks overwhelm any difficulty-induced variations in exam results that we can detect.

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 Kremer, 2010), parents (Bergbauer, Hanushek, and Woessmann 2021) 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 also reward the wrong things in classrooms (Woessmann 2018).

This paper also prompts some caution around using results on such exams for high stakes decisions and raises questions for policymakers about how to monitor progress in national education systems. A substantial empirical literature, mostly from the United States, shows that political pressures can quickly compromise performance measures (Barlevy and Neal 2012; Campbell 1976). Our findings suggest that inconsistent exam standards may be a source of substantial distortion elsewhere. This aligns with evidence from India and Pakistan which finds that a combination of poor-quality examination instruments and widespread cheating lead to spurious policy inference and inefficient government action (Burdett 2017; Singh 2020; Malik, Sarwar and Imran 2017). It also accords with assessment policy in the United Kingdom, where
a national reference test has been introduced to specifically monitor changes in standards over time (Burge and Benson 2020).

The remainder of the paper proceeds as follows. Section II. introduces our analytical approach. Section III. describes test development and data collection for hybrid tests. Section IV. presents the empirical results on examination difficulty, which we use in Section V. to estimate the economic implications of unreliable exams. Section VI. discusses options for improving assessments of this type. Section VII. concludes.

II. Analytical Approach: Converting Scores to a Common Scale

One approach to investigating the causes of the enormous variation in WASSCE pass rates would be to contrast changes in school and student ‘inputs’ with changes in achievement. However, several variables have not been measured consistently since 2011, and even where data have been obtained, research 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.
A. 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 can calculate the difficulty of each item and, collectively, the average difficulty of each examination year. We estimate the following relationship:

\[ Y_{ijkt} = \delta_j + \chi_k + \beta Year_t + \epsilon_{ijkt} \]

(Equation 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 \( \delta_j \) to control for differences in ability and since each booklet contains a unique set of items, we include booklet fixed effects \( \chi_k \) 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 $\epsilon_{ijkt}$ 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 exam difficulty changes. 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).

**B. 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.

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

(Equation 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 $\theta_j$ 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 Generalized Partial Credit Model (GPCM) is used for all items that are scored in more than two ordered categories. An item scored $0, 1, \ldots, m$ 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 $\theta_j$ states that the probability of getting a score $x$ on item $i$ denoted by $P(y_i = x|\theta_j)$ is:

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

(Equation 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 $\geq 3$ to their nearest (higher) even number. The selected IRT models are fit to the data via maximum likelihood estimation, returning all item and threshold parameters.

C. 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. The Online Appendix provides evidence from an item fit analysis that assesses the adequacy of the functional form of the ICC implied by the chosen model (AERA 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, $\theta$, 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 ability (see Section III.B.).

In applying IRT models, an assumption of local independence is also made, which means that after accounting for 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.

III. Test Development and Data Collection

A. The Structure of The WASSCE and How Grades Are Used

The WASSCE is administered to school candidates in the third and final year of senior high school. Students sit tests in four ‘Core’ subjects: mathematics and English language, which are the focus of this study, along with integrated science and social studies. Each candidate will also sit for up to five elective subjects, related to their program of study in school. For each subject there are several parts to the final exam, assessed using written tests, with other methods used sparingly. Core subjects include two written papers each. English also includes an oral assessment and integrated science a test of practical work, where facilities permit, or a third written paper. Finally, school-based continuous assessment marks contribute to student scores, which are discussed further in Section IV.E.

One of nine grades may be awarded for each WASSCE subject, from an A1 to an F9, based on a candidate’s overall score. A lower numeric score represents higher performance, with A1 considered ‘Excellent’ performance and an F9 a failing grade. Anyone scoring E8 or better will pass an individual subject. A credit-pass is awarded for grades A1 to C6.

Employers and universities require high performance in core WASSCE subjects for student selection. At the time of writing, for example, the University of Ghana requires a credit-pass (i.e., grades A1 to C6) in both English and mathematics and a mix of pass (D7 and E8) and
credit-pass (A1 to C6) awards in four other subjects. The University of Cape Coast—Ghana’s highest-ranked university—requires a credit-pass in six subjects, three of which must be core subjects. Entry to teacher training, a major public sector employer, requires students to pass six WASSCE subjects, including English and mathematics, with at least three of these credit-passes.

B. Constructing an Item Bank from Past Papers

For the purposes of this study, we develop an item bank containing WASSCE items from exams 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 (Ministry of Education 2018). We include everything from the exam that can be administered on paper and 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 exams SC4021 and SC4022 and English 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 contains five constructed-response items, each worth 8 marks, and Part II contains 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 reproduced all original examination items, word for word. 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 exam content for English and 100 percent for mathematics, affecting how we analyze and present results in Section IV.

The bank can be used to provide a first impression of exam comparability, which we look at here for mathematics. WAEC’s exam specification groups mathematics items into seven content domains and five cognitive domains. Using these categories, in Figure 1.A. we map mathematics 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 prioritized 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 1.B.). Each exam emphasizes 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. Instead, in these years and most prominent in 2017,
there is a shift towards items requiring students to ‘Analyze’, which is, ex-ante, a more difficult cognitive domain.

C. 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. 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.

D. 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 (Table 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 the 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 2022). 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, Mfantsimans Municipal, Accra Metro, Tema Metro, Ellembele, Sekondi Takoradi Metro.

E. 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 the Trends in International Mathematics and Science Study (TIMSS) (Foy and Yin 2016). 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 mathematics tests. All marks for affected items or sub-items have been omitted from the analysis (see Table 3).

IV. 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 achieving a pass (i.e., grades D7 and E8) or credit-pass (grades A1 to C6) from year to year. Third, with exam difficulty estimates, we can 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 mathematics when making comparisons with historical results.

A. Average Examination Difficulty

Across subjects and years, we identify large and statistically significant differences in average test difficulty. Table 2 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 III.B.), which aligns with the dip in pass grades shown in Figure 2. Converting estimated item difficulties into total marks would align with this, returning the highest raw scores from 2011-2013.

We see much larger differences in average item difficulty from year to year in mathematics. 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 (Table 3), 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 III.B.).
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.

**B. Pass and Credit-Pass Rates**

A full distribution of scores is required to evaluate the proportion of candidates reaching each grade boundary. 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. The probability of correctly endorsing each response category is calculated for each candidate 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 assume that the 3 percent of marks that were previously omitted (Table 3) have the same average difficulty as other items in that examination year.

After WASSCE scripts 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 receive pass and credit-pass grades each year. We set grade boundary marks by pooling pass and credit-pass 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.

We can also estimate a signal-to-noise ratio in individual WASSCE scores. Recall that we have a single group of 1,000 normally distributed candidate abilities, and that for each candidate we have a total expected score on each test. We estimate a within-candidate correlation of test scores of 0.83, implying a signal-to-noise ratio of 4.9. We will return to this later, when discussing the economic implications of an unreliable WASSCE.

C. New Evidence on Past Achievement Trends

Figure 2 presents WAEC results from Ghana alongside our difficulty-based estimates. Four features stand out. First, we track the early increase in pass and credit-pass grades awarded 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-pass awards, among all passes, in the years before 2015. Fourth, 2017 shows a large difference between WAEC results and our estimates.
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-pass awards falls by 71 percent. We split the historical variation in grade awards into two parts. One part depends on test difficulty, and another represents cohort ‘quality’, which may be used 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 3.A.). At the pass level performance rises to 2013, then returns to roughly 2011 levels before an uptick in 2019. At the credit-pass 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 mathematics 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 mathematics 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 3.B.).

From this decomposition, we estimate the number of ‘excess’ failures and ‘excess’ passes each year (Table 4). 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 mathematics WASSCE led to excess failures of a similar magnitude or, equivalently, between 15 and 35 percent of all students that failed the mathematics 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 mathematics test in 2015 would have passed in 2019.
D. Can Test Difficulty Explain Performance in Other WASSCE Countries?

The WASSCE is administered outside Ghana, giving us an opportunity to put our results in a broader context. For the period under study, tests were administered concurrently in each country, with grade award decisions made by a committee that represents all WAEC members, using inputs from each country. We have obtained administrative data on exam grades from Nigeria, The Gambia and Sierra Leone, for the core mathematics exam from 2011 to 2019. Using these, we can test if our basic approach, based on fieldwork in Ghana, can explain exam fluctuations not just in Ghana but in other countries as well.

Both the number of candidates and their performance varies enormously between WASSCE countries. In 2019, Nigeria supplied 1.55 million candidates, Ghana 0.35 million, Sierra Leone 0.12 million, Liberia 0.04 million, and The Gambia 0.01 million. From 2011 to 2019, Ghana’s average mathematics pass rate stood at 72 percent. In contrast, Nigeria, The Gambia, and Sierra Leone had pass rates of 85 percent, 22 percent, and 40 percent, respectively. As in Ghana, the proportion of students meeting the threshold for a pass or credit-pass also fluctuates over time in other countries as well, from 71 to 93 percent in Nigeria, from 13 to 40 percent in The Gambia, and from 14 to 74 percent in Sierra Leone (Figure 4).

Our basic approach helps to explain fluctuations in exam pass rates in Nigeria and The Gambia, particularly since 2014. Recall that test item difficulty is the only factor that drives variation in our simulated pass rates.

On the face of it, Nigeria’s performance trend is not easy to reconcile with test difficulty swings, unless you divide the results into two periods (Figure 5). For the second half of the decade, 2014 to 2019, we can do a good job of explaining variation in performance with test difficulty. Credit-pass grades varied from 57 to 86 percent, and we can explain 62 percent of
this variation with differences in test difficulty. We can also explain 53 percent of the variation in pass rates, despite there being much smaller changes in pass rates in Nigeria than other countries (this is what we would expect with a pass cutoff a long way from the center of the student score distribution (Ho 2008).

However, the rapid improvement in official WASSCE exam performance in Nigeria from 2011 to 2014 remains a mystery. Three possibilities are worth discussing. First, given that more than half of the change occurs from 2011 to 2012 it may be driven by a particularly low-ability 2011 cohort. However, that idea is not well supported by levels of performance in other WASSCE subjects in the same years. Second, it might represent something much worse than test-difficulty swings, such as widespread malpractice. There are numerous reports of increasing malpractice from the period, including the first reports of “miracle examination centres” (Agwu et al. 2022). Although several WAEC initiatives have been introduced to curb malpractice (Anzene 2014), there are no data available to explore this possibility further (i.e., each candidate’s response pattern, or school-based continuous assessment inputs).

A third possibility might represent something much better—i.e., genuine improvements in candidate ability. We can contrast changes in the WASSCE with results from the National Examination Council of Nigeria (NECO) for these years. NECO administers a Senior Secondary Certificate Examination at the end of secondary school in Nigeria and students may choose which examination to take, with many opting to sit both the NECO and the WASSCE. NECO results for mathematics were reported as a “mass failure” in 2011 (Vanguard 2011) with only 26 percent of candidates achieving a credit-pass. This figure rose to 50 percent in 2012, 65 percent in 2013 and 69 percent in 2014. Granted, a similar analysis might produce similar questions about the integrity of the NECO exam, which is beyond the scope of this paper. For now, we interpret the parallel improvements in scores on these two independent exams as weak
(but still inconclusive) evidence against cheating as an explanation, and in favor of genuine increases in pupil performance in the 2011-2014 period.

For The Gambia, our simulated performance tracks real grade awards over the period, with our approach helping to explain up to 28 percent of variation in pass and credit-pass awards over time. The year 2017 stands out as an anomaly in all countries, with our analysis indicating a relatively difficult test, in contrast with the high performance reported by each country. One possible explanation for this inconsistency in all countries is a central adjustment to grade boundaries, affecting all participating countries. We do not have data from WAEC’s standard fixing and grade award meetings, but a central change to the cut scores in 2017, in reaction to the very low marks obtained by students on this particularly difficult test, would be consistent with this observation in all countries, and with our hypothesis that central adjustments to test difficulty—or how test difficulty translates into grades awarded—influences WASSCE results.

Finally, Sierra Leone presents a totally different pattern of results. We do not carry out a similar analysis for Sierra Leone, as there is good reason to believe its exam performance over our study period was heavily influenced by country-specific shocks. First, the Ebola pandemic closed schools in Sierra Leone and cancelled WASSCE participation in 2013, with impacts onto candidature and preparation in the following years. From an entry of 58,000 in 2012, WASSCE sitters fell by more than half, to 28,000 in 2014, then further to 24,000 in 2015 before recovering in 2016-2018. Second, Sierra Leone changed the structure of secondary school for the cohort entering in 2016, reducing the number of senior secondary grades from 4 to 3 and doubling the exam cohort arriving in 2019. Third, a change of rules introduced under the Free Quality School Education policy of 2018 led to schools “entering candidates that were not regular students of their schools” (MBSSSE 2019). Altogether exam candidature jumped more than three-fold, to 115,000 in 2019 with the lowest performance on record since 2011. These
country-specific swings in candidate selection and preparation overwhelm any difficulty-induced variation in pass rates that we can identify.

E. Limitations

Our results cannot explain all the change in grades awarded over time in Ghana, much less across Anglophone West Africa. In addition to grade boundary adjustments (Section IV.B.) there are at least five 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 mathematics 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 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 3.A.) result from genuine differences in student preparation. There have been several changes
to secondary education in Ghana over the past decade, which have sought to increase access and improve quality in low-performing schools. It is, however, beyond the scope of this paper to investigate or evaluate the impacts of Ghana’s many education reforms on student outcomes.

(5) The students in Ghana used to generate item parameters do not perform at the same levels as their peers in Nigeria or The Gambia. While this does not introduce any bias, it may limit our ability to convert scores towards the top and bottom of the skill distribution and limit our ability to explain variation in grade awards in countries other than Ghana (Kolen and Brennan 2014).

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 in Ghana but perhaps a larger contribution elsewhere.

V. Economic Implications of Unreliable Exams

Besides their implications for fairness or student incentives, large fluctuations in exam difficulty may undermine productivity through the misallocation of talent. In this section, we sketch a simple framework to think about what our results mean for worker productivity, which we apply to new estimates of the return to observable skills in the Ghanaian labor market. Our calculations suggest that improvements in the comparability of the WASSCE exam would increase the Mincerian return to skills by about 6 percentage points among workers completing secondary school.
A. Framework

As a benchmark, assume that if employers had perfect information about worker skills, the equilibrium wage rate, \( w \), would entail a return to skills, \( y \), in line with a typical Mincerian model:

\[
\ln(w) = b_0 + b_1 y
\]

(Equation 4)

Now suppose employers are unable to observe \( y \), and instead judge workers’ skill based on their WASSCE performance, which is an imperfect signal of their true skill such that \( y_0 = y + e \), where \( e \) represents random noise. Employers’ best estimate of a worker’s true skill will be \( \mathbb{E}[y|y'] = yy' \), where:

\[
y = \frac{\sigma^2_y}{\sigma^2_y + \sigma^2_e} < 1
\]

(Equation 5)

The skill premium on observed WASSCE scores will be \( b_1 y \), i.e., less than the return to skills under perfect information. Unfortunately, we know of no labor market survey in Ghana with detailed data on WASSCE performance.

What we do observe in survey data is a separate, independent proxy of a worker’s true skill, \( y'' = y + u \), where we assume \( \text{cov}(e,u) = 0 \). Specifically, we rely on Ghana’s Skills Towards Employability and Productivity (STEP) surveys, which gather information from a representative sample of urban adults between the ages of 15 and 64. The survey includes a detailed assessment of literacy skills, with scores bench-marked to the OECD’s Program for
the International Assessment of Adult Competencies (PIAAC) assessment. Participants also complete a background questionnaire that gathers information about labor-market status, earnings, education, and demographic characteristics.

PIAAC scores provide a fine-grained measure of workers skills which is unlikely to be observed by employers. For convenience, denote the linear projection of one skill measure (WASSCE) onto the other (PIAAC) by \( y' = \phi y'' + \nu \), where again, \( y' \) is observed by employers but not in our survey data, and vice versa for \( y'' \). Using data on wages and independent test scores for individual workers indexed by \( i \), we can estimate:

\[
\ln(w_i) = \delta_0 + \delta_1 y''_i + \epsilon_i
\]

(Equation 6)

Where the return to skills observed in the data should be a fraction of the return employers actually pay for WASSCE scores, which is itself lower than the perfect-information benchmark, i.e., \( \delta_1 = \beta_1 \gamma \phi \).

Now we can consider the thought experiment where the comparability of the WASSCE is improved such that \( e = 0 \) for all individuals. The observed return to our independent measure of actual skills, \( \delta_1 \), would increase from \( \beta_1 \gamma \phi \) to \( \beta_1 \phi \), i.e., returns to real skills would be scaled up in inverse proportion to the signal-to-noise ratio of the original, imperfect WASSCE scores.

While this framework relies on fairly strong simplifying assumptions about wage determination, it allows us to put some empirical numbers on the wage and productivity consequences of exam comparability based on actual labor market data.
B. Returns to Measured Skills in Ghana’s Labor Market

To obtain an estimate of $\delta_1$, we regress labor market earnings on PIAAC scores from the STEP survey. Our empirical specification is a standard extension of the basic Mincer equation, using measured cognitive skills as a proxy of human capital, rather than (only) years of schooling.

$$\ln(w_i) = \delta_0 + \delta_1 Skills_i + \delta_2 Age_i + \delta_3 Age_i^2 + \delta_4 Female_i + \epsilon_i$$

(Equation 7)

Here $\ln(w_i)$ denotes Log hourly earnings, $Skills$ is our standardized measure of literacy skill, with a mean of zero and a standard deviation of one, $Age$ and $Age^2$ capture potential labor market experience, $Female$ is a gender indicator, and $\epsilon$ a stochastic error. We are interested in $\delta_1$, the earnings gradient associated with measured skill.

We limit our analytical sample to adults between the ages of 25 and 64 who have completed secondary school and we exclude those currently attending any educational program. We retain only waged and salaried workers and those who are self-employed. Employers were excluded from the self-employed group to avoid measurement error in reported earnings. We winsorize the top and bottom one percent of the wage distribution to limit the influence of outliers.

Part of the return to cognitive skills comes through continuation in school. Entry into higher education and entry into public sector employment both depend on a candidate passing the WASSCE. To show the empirical relevance of school attainment (university or other tertiary) downstream of completing secondary school, we add years of schooling in a later specification of this model. In a final specification we add an indicator for public sector employment.
Among this sample of Ghanaian secondary school completers, we observe significant positive returns—both in terms of earnings and the probability of holding a wage job—to additional years of schooling, and to test scores measured on independent and high-quality learning assessments.

**C. Wage and Productivity Consequences of Unreliable Exams**

We can now combine parameters from our test and labor market data to simulate the wage consequences of improvements in the comparability of the WASSCE exam.

We take our return to skills from specification (5) in Table 5, which controls for years of schooling and public sector employment downstream of WASSCE completion. This is our empirical estimate of $\delta_1 = \beta_1 \gamma \phi = 0.238$, and implies that a one standard deviation increase in measured literacy provides a 27 percent increase in hourly earnings. We combine this with our estimate of $\gamma = 0.83$, calculated from WASSCE test data in Section IV.B.

Now we posit that WASSCE comparability can improve such that $e = 0$ and hence $\gamma$ rises to 1. In this case, $\delta_1$ would increase to 0.287, i.e., the observed wage return would rise by 6 percentage points, or 22 percent.

**VI. Discussion: Policy Options to Improve Exam Comparability**

Inconsistent exam standards can be a source of substantial distortion in countries that follow the same assessment model as used in the WASSCE. Kellaghan and Greaney (2019) identify a psychometric assessment method, which places emphasis on item discrimination and test reliability. This approach is the basis for the Scholastic Aptitude Test (SAT) and large-scale international assessments such as the Programme for International Student Assessment (PISA) and TIMSS. They contrast this with a curriculum-based assessment method which is primarily
concerned with assessing whether a student has acquired the knowledge and skills listed in the curriculum. This approach places a strong emphasis on validity and results tend to be given in the form of grades, with no requirement that scores are normally distributed. It underpins the WASSCE and dominates high stakes testing in most education systems.

Furuta (2021) shows, for a sample of 138 countries in 2010, that almost 90 percent of countries use high-stakes curriculum-based exams at the senior-secondary level of the educational process. Among 30 nations that participate in the Programme for the International Assessment of Adult Competencies, 23 hold central exit exams at the end of high school, with the USA a notable exception (Leschnig, Schwerdt, and Zigova 2022). A new database of high-stakes exams (Rossiter and Konate 2022) also shows that 40 of Africa’s 55 countries hold this type of curriculum-based exam at the end of primary school and 45 of these at the end of lower secondary school. As is the case for the WASSCE, these exams are used to award credentials, to filter children into the next level of education and/or to determine the type of school that they can attend. Countries often combine approaches to maintain test comparability over time and minimize the types of distortion we have identified for the WASSCE.

Technical fixes involving more formulaic approaches to maintaining exam standards could partially resolve the problems identified in this study. Expert judgment is always going to be needed to maintain consistent standards. But where systems depend too much on expert judgment—as might be the case for the WASSCE—they may compensate for this by strengthening the role of statistical methods to link assessments over time and use expert judgment in a confirmatory capacity. Exam boards might even use common anchor items and Item Response Theory to make a direct link between tests (Burdett et al. 2013; Pointer 2014). But this can also give a false sense of precision.
A more promising approach might be to use an external reference test directly to set standards. Used alongside the current WASSCE, changes in performance on the reference test could inform adjustments to grade boundaries each year. Hong Kong appears to have been the first country to follow this approach, now adopted in the UK (Burdett et al. 2013). There, the National Reference Test (NRT) was introduced in 2017. Each year a random sample of students takes a test in mathematics or English shortly before they sit their secondary school exams. Results are analyzed at the national level with no consequences for the sampled students or schools. Tests are statistically linked over time and the main benefit of this approach is that it indicates if the percentage of students reaching important benchmarks changes from year to year. It is these changes, if any, that will be considered when student exam results are awarded.

However, these technical fixes fail to address the underlying challenge of using a single high stakes assessment for both screening pupils and maintaining educational standards. As Barlevy and Neal (2012) have argued, it may be preferable for WAEC members to develop separate assessment systems that are designed specifically for each measurement task.

This could take on many forms, but one option is for a government to monitor system performance using a national assessment on a sample basis. Test items can be statistically linked as the stakes are low for students and teachers, and results capture how overall performance is changing. With a sample-based assessment in place, officials can be freed from the need to place WASSCE (or other exam) results on a common scale. High stakes exams need not follow a common format and can be less predictable from one year to the next, reducing opportunities for ‘gaming’, such as through teaching to the test. Grade awards can simply reflect the ordinal information contained in assessment results which is likely to be a sufficient measure for allocating university admissions and jobs. It may also be a better metric with which to hold schools to account than grade awards that can vary arbitrarily from one year to the next.
VII. 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 standardized 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 mathematics grades awarded in Ghana. As a result, education officials and civil society monitor progress using unreliable performance metrics.

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 Kremer 2021). 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).

As would be expected of a regionally administered test, exam difficulty alone also explains much of the variation in grade awards in Nigeria and The Gambia. However, the rapid improvement in official WASSCE exam performance in Nigeria from 2011 to 2014 remains something of a mystery. Despite reports of widespread examination malpractice, we interpret the improvements in scores on the WASSCE, and on parallel exams, as weak evidence in favor of genuine increases in candidate skills in the 2011-2014 period. In the context of declining
education quality in the developing world (Le Nestour, Moscoviz, and Sandefur 2022), this period of rapid performance improvement in Nigeria warrants further investigation.

The economic implications of the fluctuations in exam difficulty that we identify are potentially large. In a simple framework where employers use WASSCE scores to infer workers’ true productivity, unreliable exams will reduce the skill premium. Combining parameters from our main analysis with Ghanaian survey data on the return skills measured through high-quality, independent learning assessments, we calculate that removing spurious variation in WASSCE exam difficulty would increase the observed wage return to skills by 6 percentage points among workers completing secondary school.

Our analysis is not without limitations. We cover most of the information from terminal exams but cannot account for continuous assessment and other information that WAEC uses to award grades. 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 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.

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


Figures

Panel A: Marks Categorized by Year and Content Domain

- Stats / Prob.
- Trigonometry
- Coordinate Geo.
- Plane Geo.
- Mensuration
- Algebra
- Number


* Multiple-Choice Items ○ Constructed-Responses Items

Panel B: Marks Categorized by Year and Cognitive Domain

- Evaluate
- Analyze
- Apply
- Comprehend
- Recall


* Multiple-Choice Items ○ Constructed-Responses Items

Figure 1

Item Domains & Marks Available in WASSCE Mathematics Exams From 2011-19.

Note: Figures map test items by year and content domain (Panel A) and by year and cognitive domain (Panel B). 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. In Panel B, markers are slightly offset in Comprehend and Recall domains, where equal marks are available for each item type.
Figure 2

Actual WASSCE Pass Rates for Ghana Compared with Difficulty-Induced Pass Rates in Our Study.

Note: Panel A shows results for the population of WASSCE takers each year. Panel B shows the pass rates among our single sample of students. ‘Credit or better’ is a WASSCE grade A1 to C6. ‘Pass’ is a WASSCE grade D7 or E8.
Panel A: Difficulty-Adjusted Performance Trend in Mathematics

Panel B: Variation in Pass Rates Attributed to Test Difficulty and to The Cohort

Figure 3

Deconstructing Historical Trends in Ghana’s WASSCE Performance.

Note: We separate the differences in pass rates into two parts: one can be explained by test difficulty, the other cannot. For convenience we refer to this latter part as the Cohort, indicating a genuine change in cohort performance. This is a residual, we cannot, with available data, directly test any hypothesis about cohort competency. All values are shown relative to an average difficulty test, based on pooled data for 2011-2019. Our estimates of test difficulty do not provide a reasonable explanation for the sharp increase in pass grades awarded in 2017, so we omit that year from this figure.
Figure 4


Note: figure shows administrative data on student achievement in three of the five WASSCE member countries, Nigeria, The Gambia and Sierra Leone. Data are obtained from WAEC (Nigeria) and the Ministry of Education (The Gambia and Sierra Leone) in each country. Student results are shown for pass and credit-pass thresholds, and these represent the same levels of performance in every country. There is no result shown for Sierra Leone in 2013 because the Ebola pandemic prevented the country from fielding any WASSCE candidate in that year. We were unable to obtain data for Liberia for the period of this study.
Figure 5

Simulated WASSCE Performance Holding Exam Difficulty Constant.

Note: figure shows the results of our performance simulation for Nigeria and The Gambia. Each panel shows our estimates of pass and credit-pass rates, with test difficulty the only source of variation over time. For each simulation we use the same item parameters and grade boundaries as generated in our main analysis for Ghana.
Table 1
Student and School Demographics

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Our Sample</td>
<td>All of Ghana</td>
</tr>
<tr>
<td><strong>Panel A: Student Characteristics</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td>17.76</td>
<td>19.32</td>
</tr>
<tr>
<td>Male</td>
<td>38.42</td>
<td>50.02</td>
</tr>
<tr>
<td><strong>Panel B: School Characteristics</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Category A</td>
<td>9</td>
<td>55</td>
</tr>
<tr>
<td>Category B</td>
<td>11</td>
<td>220</td>
</tr>
<tr>
<td>Category C</td>
<td>9</td>
<td>370</td>
</tr>
</tbody>
</table>

Note: values for All of Ghana are based on 2018 WASSCE data.
Table 2

Average Marks Awarded Per Item, for Each Subject and Examination Year (Our Study)

<table>
<thead>
<tr>
<th></th>
<th>English</th>
<th></th>
<th>Mathematics</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(1)</td>
<td>(2)</td>
<td>(3)</td>
<td>(4)</td>
<td>(5)</td>
<td>(6)</td>
</tr>
<tr>
<td>2011</td>
<td>-0.013***</td>
<td>-0.011***</td>
<td>-0.011***</td>
<td>0.095***</td>
<td>0.082***</td>
<td>0.082***</td>
</tr>
<tr>
<td></td>
<td>(0.004)</td>
<td>(0.004)</td>
<td>(0.004)</td>
<td>(0.017)</td>
<td>(0.017)</td>
<td>(0.017)</td>
</tr>
<tr>
<td>2012</td>
<td>-0.050***</td>
<td>-0.047***</td>
<td>-0.047***</td>
<td>0.137***</td>
<td>0.133***</td>
<td>0.136***</td>
</tr>
<tr>
<td></td>
<td>(0.004)</td>
<td>(0.004)</td>
<td>(0.004)</td>
<td>(0.017)</td>
<td>(0.017)</td>
<td>(0.016)</td>
</tr>
<tr>
<td>2013</td>
<td>-0.025***</td>
<td>-0.031***</td>
<td>-0.031***</td>
<td>0.078***</td>
<td>0.071***</td>
<td>0.070***</td>
</tr>
<tr>
<td></td>
<td>(0.004)</td>
<td>(0.004)</td>
<td>(0.004)</td>
<td>(0.017)</td>
<td>(0.017)</td>
<td>(0.016)</td>
</tr>
<tr>
<td>2014</td>
<td>-0.045***</td>
<td>-0.048***</td>
<td>-0.048***</td>
<td>0.120***</td>
<td>0.115***</td>
<td>0.112***</td>
</tr>
<tr>
<td></td>
<td>(0.004)</td>
<td>(0.005)</td>
<td>(0.004)</td>
<td>(0.017)</td>
<td>(0.017)</td>
<td>(0.016)</td>
</tr>
<tr>
<td>2016</td>
<td>0.022***</td>
<td>0.022***</td>
<td>0.023***</td>
<td>0.049***</td>
<td>0.036**</td>
<td>0.036**</td>
</tr>
<tr>
<td></td>
<td>(0.004)</td>
<td>(0.005)</td>
<td>(0.004)</td>
<td>(0.017)</td>
<td>(0.017)</td>
<td>(0.016)</td>
</tr>
<tr>
<td>2017</td>
<td>-0.033***</td>
<td>-0.028***</td>
<td>-0.028***</td>
<td>-0.017</td>
<td>-0.028</td>
<td>-0.027*</td>
</tr>
<tr>
<td></td>
<td>(0.004)</td>
<td>(0.005)</td>
<td>(0.004)</td>
<td>(0.017)</td>
<td>(0.017)</td>
<td>(0.016)</td>
</tr>
<tr>
<td>2018</td>
<td>-0.049***</td>
<td>-0.045***</td>
<td>-0.045***</td>
<td>-0.011</td>
<td>-0.024</td>
<td>-0.026</td>
</tr>
<tr>
<td></td>
<td>(0.004)</td>
<td>(0.005)</td>
<td>(0.004)</td>
<td>(0.017)</td>
<td>(0.017)</td>
<td>(0.016)</td>
</tr>
<tr>
<td>2019</td>
<td>-0.044***</td>
<td>-0.050***</td>
<td>-0.050***</td>
<td>0.340***</td>
<td>0.326***</td>
<td>0.322***</td>
</tr>
<tr>
<td></td>
<td>(0.004)</td>
<td>(0.005)</td>
<td>(0.004)</td>
<td>(0.017)</td>
<td>(0.017)</td>
<td>(0.016)</td>
</tr>
<tr>
<td>Constant</td>
<td>0.541***</td>
<td>0.541***</td>
<td>0.541***</td>
<td>0.642***</td>
<td>0.651***</td>
<td>0.652***</td>
</tr>
<tr>
<td></td>
<td>(0.003)</td>
<td>(0.003)</td>
<td>(0.003)</td>
<td>(0.012)</td>
<td>(0.012)</td>
<td>(0.012)</td>
</tr>
</tbody>
</table>


Booklet Fixed Effects: No, Yes, Yes, No, Yes, Yes

Pupil Fixed Effects: 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 mathematics 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.
Table 3

Total Marks Per Year for Mathematics WASSCE

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Marks Available</td>
<td>186</td>
<td>186</td>
<td>186</td>
<td>186</td>
<td>186</td>
<td>186</td>
<td>186</td>
<td>186</td>
<td>186</td>
</tr>
<tr>
<td>of which, objective-type items</td>
<td>50</td>
<td>50</td>
<td>50</td>
<td>50</td>
<td>50</td>
<td>50</td>
<td>50</td>
<td>50</td>
<td>50</td>
</tr>
<tr>
<td>of which, subjective-type items</td>
<td>136</td>
<td>136</td>
<td>136</td>
<td>136</td>
<td>136</td>
<td>136</td>
<td>136</td>
<td>136</td>
<td>136</td>
</tr>
<tr>
<td>Marks Omitted from our analysis</td>
<td>10</td>
<td>8</td>
<td>0</td>
<td>2</td>
<td>5</td>
<td>3</td>
<td>2</td>
<td>24</td>
<td>2</td>
</tr>
<tr>
<td>Percent Omitted from our analysis</td>
<td>5</td>
<td>4</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>13</td>
<td>1</td>
</tr>
</tbody>
</table>

Note: shows Marks Available for Core Mathematics WASSCE SC4021 and SC4022. We also show the Marks Omitted and Percent (of marks available) Omitted from our analysis. Of the 450 objective-type items in the bank, 9 include an error in the stem which was introduced when constructing the item bank. Marks for these items are omitted from the analysis. The 117 constructed response items contain 468 sub-parts. Ten sub-parts include an error introduced when constructing the item bank. These errors affect 47 marks out of 1,224 total. All marks for affected items or sub-items have been omitted from the analysis.
## Table 4

**Difficulty-Induced Excess Failures in the Mathematics WASSCE in Ghana**

<table>
<thead>
<tr>
<th>Exam Year</th>
<th>Exam Candidates</th>
<th>Reported Fail Rate (%)</th>
<th>Excess Fails</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Candidates</td>
</tr>
<tr>
<td>2011</td>
<td>147,092</td>
<td>23</td>
<td>-11,326</td>
</tr>
<tr>
<td>2012</td>
<td>172,913</td>
<td>18</td>
<td>-21,095</td>
</tr>
<tr>
<td>2013</td>
<td>406,001</td>
<td>28</td>
<td>26,796</td>
</tr>
<tr>
<td>2014</td>
<td>237,683</td>
<td>32</td>
<td>18,539</td>
</tr>
<tr>
<td>2015</td>
<td>262,913</td>
<td>46</td>
<td>43,381</td>
</tr>
<tr>
<td>2016</td>
<td>268,187</td>
<td>38</td>
<td>21,455</td>
</tr>
<tr>
<td>2017*</td>
<td>286,544</td>
<td>20</td>
<td></td>
</tr>
<tr>
<td>2018</td>
<td>312,486</td>
<td>32</td>
<td>15,937</td>
</tr>
<tr>
<td>2019</td>
<td>342,529</td>
<td>14</td>
<td>-34,938</td>
</tr>
<tr>
<td></td>
<td>2,436,348</td>
<td>28</td>
<td></td>
</tr>
</tbody>
</table>

Note: Exam Candidates and Reported Fail Rate reflect real test outcomes, reported by WAEC for each year. Reported Fail Rate is the proportion of all candidates who received an F9 grade. Excess Fails columns show the number (Candidates) and proportion (% Candidates) of students who we estimate failed the mathematics WASSCE in that year, because of changing test difficulty. All values are relative to a test of ‘average difficulty’, based on pooled data from 2011-2019. Positive (negative) values indicate higher (lower) rates of failure in that year than would have been expected if test difficulty had not changed over the period. *We cannot explain the sharp increase in ‘Pass’ grades awarded in 2017 based on our data on test difficulty, so we omit any estimation of excess fails for this year. Source: Exam Candidates and Historical Fail Rate from WAEC; Excess Fails based on authors’ calculations.
## Table 5

The Return to Skill in Ghana’s Labor Market

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Wage Emp.</td>
<td>Wage Emp.</td>
<td>Earnings</td>
<td>Earnings</td>
<td>Earnings</td>
</tr>
<tr>
<td>Literacy</td>
<td>0.138***</td>
<td>0.039</td>
<td>0.505***</td>
<td>0.267**</td>
<td>0.238*</td>
</tr>
<tr>
<td></td>
<td>(0.039)</td>
<td>(0.042)</td>
<td>(0.120)</td>
<td>(0.132)</td>
<td>(0.129)</td>
</tr>
<tr>
<td>Age</td>
<td>0.026</td>
<td>0.014</td>
<td>0.134**</td>
<td>0.112**</td>
<td>0.098*</td>
</tr>
<tr>
<td></td>
<td>(0.021)</td>
<td>(0.020)</td>
<td>(0.054)</td>
<td>(0.053)</td>
<td>(0.052)</td>
</tr>
<tr>
<td>Age Squared</td>
<td>-0.000</td>
<td>-0.000</td>
<td>-0.001**</td>
<td>-0.001**</td>
<td>-0.001*</td>
</tr>
<tr>
<td></td>
<td>(0.000)</td>
<td>(0.000)</td>
<td>(0.001)</td>
<td>(0.001)</td>
<td>(0.001)</td>
</tr>
<tr>
<td>Female</td>
<td>-0.201***</td>
<td>-0.203***</td>
<td>-0.030</td>
<td>-0.055</td>
<td>-0.013</td>
</tr>
<tr>
<td></td>
<td>(0.053)</td>
<td>(0.051)</td>
<td>(0.143)</td>
<td>(0.137)</td>
<td>(0.137)</td>
</tr>
<tr>
<td>Years Schooling</td>
<td></td>
<td></td>
<td>0.067***</td>
<td>0.150***</td>
<td>0.120***</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(0.013)</td>
<td>(0.038)</td>
<td>(0.040)</td>
</tr>
<tr>
<td>Public Sector Employee</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.474***</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(0.148)</td>
</tr>
</tbody>
</table>

| Observations     | 346    | 346    | 270    | 270    | 270    |

Note: OLS regressions. Dependent variable: Wage Emp. = has wage employment, Earnings = log hourly earnings. Sample: respondents aged 25 to 64, who have completed secondary school and are no longer in education. Columns 1-2 include individuals not working, while columns 3-5 are restricted to individuals with positive wage or self-employment earnings. Employers’ earnings are omitted from the self-employed group to avoid measurement error in reported earnings. The top and bottom one percent of earnings are winsorised. * \( p < 0.10 \), ** \( p < 0.05 \), *** \( p < 0.01 \). Robust standard errors are shown in parentheses. Source: World Bank STEP survey for Ghana, 2013.
Endnotes

1 http://www.waecgh.org/about-us

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

3 Although 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 mathematics).

4 Other potential sources of variation in grades awarded are discussed in Section IV.E.

5 We follow the same approach as in our main analysis, explained in Section IV.B. We carry over parameters for all mathematics items, along with the marks required to reach pass and credit-pass thresholds, to simulate achievement for Nigeria and The Gambia. For each country, we generate a single set of students, of normally distributed ability, with mean and standard deviation computed to minimise the difference between average actual and implied results over the period.

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