# Blindsided at the Start: Small Shocks and High-Stakes Performance

Yi Chen, Sikun Dou, Hongbin Li, Roujing Wu (JMP)

This version: September 16, 2026

**Link to Newest Version**

**Abstract**

This paper examines whether a shock with limited direct stakes at the start of a high-stakes assessment affects outcomes beyond those stakes. Using National College Entrance Examination in China, the analysis links administrative records for 13.6 million students to a large-language-model-based difficulty scale covering 484 questions from 22 mathematics exam sets. The empirical design exploits variation across administratively assigned exam sets while conditioning on aggregate exam difficulty. A one-standard-deviation increase in first-question difficulty lowers mathematics scores by 0.252 standard deviations, which is equivalent to 6.4 points. Since the question is worth at most five points and can be revisited, its direct score value cannot account for the full loss. Students from good high schools experience losses that are 0.082 standard deviations smaller. When applications are fixed before the exam, the same increase raises their STEM University admission probability relative to other students by 1.4 percentage points, corresponding to an estimated reallocation of roughly 7,650 STEM seats toward them. These patterns align with

Yi Chen: ShanghaiTech University. Sikun Dou: Singapore Management University. Hongbin Li: Stanford University. Roujing Wu: University of Bologna. We would like to thank Benjamin Arold, Davide Cantoni, Hanming Fang, Henning Hermes, Vincenzo Scrutinio, and Weilong Zhang for suggestions that significantly improved the paper. This paper was presented at the University of Bologna, the University of Cambridge, and Ludwig Maximilian University of Munich. Chen acknowledges financial support from the National Science Foundation of China (72422019). All remaining errors are our own.

expectation-based salience, where an unexpected opening shock operates through persistent cognitive interference rather than time depletion to disrupt performance across the assessment. An adverse start can therefore affect both performance and the allocation of educational opportunities.

**Keywords:** opening shocks, high-stakes testing, exam design, salience, educational inequality, college admissions

**JEL Classification:** I21, I24, D91

# 1. Introduction

Some small shocks end where they occur. Others continue to shape what follows. An adverse opening in a high-stakes task provides a particularly revealing case. Its direct stakes can be small and precisely bounded, yet it arrives against expectations formed before the task begins, when most of the task remains. This paper studies whether an unexpected opening shock has consequences beyond its direct stakes, who is better able to absorb it, and what mechanisms can account for the additional effect.

Identifying such amplification requires separating an adverse opening from both the difficulty of the task as a whole and the characteristics of those who face it. A difficult opening may simply be one feature of a generally difficult task, while those exposed to it may differ in ability or preparation. Moreover, even when exposure is plausibly exogenous, an effect on total performance is difficult to interpret unless the maximum score loss the opening can cause directly is known. A credible design therefore requires variation in opening difficulty that is not driven by those who face it, a transparent upper bound on the opening's direct stakes, and a way to hold fixed overall task difficulty.

China's National College Entrance Examination (NCEE, or *Gaokao*) provides these elements in an unusually high-stakes setting. Performance on the examination largely determines whether and where a student attends college, with long-run consequences for education and earnings (Jia and Li 2021; Li et al. 2012). Students take the mathematics exam designated for their province and year and do not choose among exam forms. The mathematics exam also follows a stable convention: it begins with an accessible question. Students repeatedly encounter this pattern in classroom exercises, mock examinations, and past exams, giving them a clear reason to expect an easy start. The opening question carries limited direct stakes: it is worth only four or five points out of 150, and students may skip it and return later. Occasionally, however, an exam set departs from this convention and begins with an unusually difficult question. The

first question thus plays two roles: it contributes directly to the score and provides the student's earliest signal about the realized exam. Because its maximum direct score loss is transparently bounded, evaluating this opening shock while conditioning on overall exam difficulty allows us to isolate behavioral amplification from both direct mechanical point loss and overall exam difficulty.

Studying this question requires a measure of question difficulty constructed without using realized performance in the administrative records. These records report only subject-level scores, not responses to individual questions, making conventional item-response methods infeasible. We therefore digitize 484 questions from 22 mathematics exam sets between 1999 and 2003 and use repeated pairwise comparisons by GPT4o, a large language model (LLM) to place them on a common difficulty scale. The resulting measure reproduces the intended within-exam difficulty gradient, closely tracks independent assessments by mathematics experts, predicts realized performance in math exam. We then link the question-level measures to administrative data covering approximately 13.6 million NCEE participants from 1999 to 2003 in science and comprehensive tracks.

Our empirical strategy exploits variation in first-question difficulty across exam sets, controlling for aggregate exam difficulty, province and year fixed effects, and individual characteristics. The assignment of exam sets to provinces was an administrative decision driven by security protocols and the staggered rollout of curricular reform, rather than by contemporaneous local conditions, and students cannot select into the exam set assigned to their province-year. Consistent with this assignment rule, first-question difficulty is not systematically related to individual, and province-year characteristics. Together, the institutional assignment rule and this balance evidence support interpreting the coefficient on first-question difficulty as the effect of an adverse opening shock on performance, rather than a proxy for unobserved differences in student composition or local conditions.

We find that a difficult first question substantially lowers performance. A one-standard-deviation increase in first-question difficulty reduces mathematics scores by 0.252 standard deviations, equivalent to 6.4 points on a 150-point exam. Since the first question is worth at most five points and students can return to it later, the maximum direct loss on that item cannot account for the full estimated decline. The same increase in first-question difficulty also lowers the aggregate NCEE score by 0.198 standard deviations. Consistent with the shock being common to all students taking a given exam, we detect little average effect on relative rank or admission outcomes in the full sample. The absence of an average effect on these relative outcomes, however, does not imply distributional neutrality if students differ in their sensitivity to the same shock.

We therefore ask whether students differ in their responses to the same opening shock. We organize the heterogeneity analysis around high-school quality as it summarizes students' prior educational environment, including student selection, peer composition, school resources, and preparation for unfamiliar problems under pressure. Students from good high schools are substantially less affected: following a one-standard-deviation increase in first-question difficulty, the estimated decline in mathematics scores is 0.082 standard deviations smaller for these students, and the differential increases under more selective definitions of school quality. Several additional results suggest that this pattern does not merely reflect general differences in academic achievement. Students from good high schools are not systematically less sensitive to aggregate exam difficulty; the differential remains positive across deciles of realized NCEE performance; and, within schools, higher- and lower-performing students respond similarly to opening difficulty. These findings indicate that the relative advantage of students from good high schools is specific to an adverse opening rather than a general advantage in taking difficult exams or a simple gradient in observed academic achievement.

To isolate the allocative consequences operating srictly through exam performance rather than endogenous application responses, we restrict attention to province-years

in which students submit their university applications prior to the NCEE. As application portfolios are fixed before the opening shock is realized and cannot subsequently be revised, admission outcomes in this subsample cannot reflect post-exam retargeting. The negative average effects on mathematics and aggregate NCEE scores remain in this sample. More importantly, the differential response by high-school quality extends from test scores to relative rank and STEM university admission. A one-standard-deviation increase in the first-question difficulty widens the STEM university admission advantage of students from good high schools by 1.4 percentage points. Among the 546,273 such students in the pre-exam sample, this magnitude is equivalent to approximately 7,650 STEM university admissions. Thus, when application behavior is held fixed, a common opening shock can change the composition of students receiving scarce educational opportunities.

We next investigate why the effect of first-question difficulty exceeds the points directly at stake. The evidence is consistent with reference-dependent salience. Since an easy first question is a well-established feature of the mathematics exam, an unusually difficult one represents a sharp departure from students' likely ex ante reference point. Three patterns support this interpretation. First, comparable difficulty predicts lower total performance when it appears in Question 1, but not when it appears in Questions 2–4 or in later blocks of the exam. Second, performance losses are larger when the first question is more difficult relative both to the average difficulty of first questions and to the aggregate difficulty of its own exam set. Consistent with the latter comparison, the estimated effect is substantially larger on below-median-difficulty exam sets, where the contrast between the first question and the remainder of the exam is sharper. Third, the relationship is nonlinear: relative to an easy first question, one of ordinary difficulty has little detectable effect, whereas a first question above the 90th percentile of the difficulty distribution lowers mathematics scores by 0.513 standard deviations. These patterns accord with theories in which stimuli that stand out relative to a reference point receive

disproportionate attention and salient early experiences influence subsequent attention and recall (Bordalo, Gennaioli, and Shleifer 2012, 2020, 2022).

A salient first question may affect performance on subsequent items through two distinct channels. The first is a time-allocation channel: students may devote more time to the first question, thereby reducing the time available for the remainder of the exam. The second is a cognitive-interference channel: an unexpected or cognitively demanding question may continue to absorb scarce attentional and working-memory resources after students move on (Deck and Jahedi 2015; Deck, Jahedi, and Sheremeta 2021). The empirical evidence is more consistent with the second channel. Calculation complexity, our proxy for the time requirements of the first question, has a small and statistically insignificant association with total mathematics performance. In contrast, a content-based index of cognitive demand is negatively and statistically significantly associated with scores. This relationship is driven by abstractness, comprehension difficulty, and innovativeness, rather than by the number of knowledge points tested. Because we do not observe students' real-time time allocation or mental states, these estimates should not be interpreted as a formal causal mediation test. They nevertheless suggest that the performance loss beyond the points directly at stake is more consistent with persistent cognitive interference than with time allocation alone.

The paper's main contribution is to show how a disturbance with limited direct stakes can be amplified within a high-stakes selection process. Existing research demonstrates that examination performance responds to transitory conditions unrelated to students' academic preparation, including pollution, temperature, physiological stress, and cognitive fatigue (Ebenstein, Lavy, and Roth 2016; Graff Zivin et al. 2020; Heissel et al. 2021; Sievertsen, Gino, and Piovesan 2016). Related research demonstrates that the ordering of easy and difficult questions affects test completion and accuracy (Anaya et al. 2022), and that variation in question order can be used to measure cognitive endurance (Reyes 2023). We document a distinct form of within-assessment amplification. Because the

first question carries only four or five points, even a complete failure to answer it can mechanically reduce the score by no more than that amount. Separately controlling for aggregate exam difficulty distinguishes the additional effect of first-question difficulty from the consequences of facing a generally difficult exam set. The estimated decline exceeding the question's maximum point value therefore indicates that its effect extends beyond the points directly assigned to that item. Linking this variation to administrative admission records further shows when a disturbance with small direct stakes can affect the allocation of scarce educational opportunities. This evidence complements Bruhn et al. (2025). Whereas they show that aggregate scores can conceal economically relevant variation across individual items, we show that difficulty arising at a single item can affect the aggregate score by more than the item's own scoring weight and thereby influence the outcome on which selection is based.

The pattern of amplification also speaks to the economics of salience. Salience theories emphasize that a stimulus can receive disproportionate attention when it departs sharply from its context or from a reference point (Bordalo, Gennaioli, and Shleifer 2012, 2020, 2022). The closest empirical precedent is Duquennois (2022), who shows that financially salient content in mathematics questions disproportionately lowers disadvantaged students' performance and generates spillovers to subsequent items. Our setting points to a complementary source of salience: the violation of an expectation established before the assessment begins. Students enter the NCEE mathematics exam expecting an accessible first question, reflecting a longstanding feature of exam design. An unusually difficult first question therefore departs sharply from their ex ante reference point before they have received other information about the realized exam. The concentration of the effect in the first question, its greater magnitude when first-question difficulty contrasts more sharply with relevant difficulty benchmarks, and its nonlinear concentration among exceptionally difficult first questions are consistent with this interpretation. Evidence further suggests that the excess loss is more closely

associated with cognitive demands than with the time required to solve the question. Since attention, time allocation, and working memory are not directly observed, these results provide suggestive evidence on the mechanism rather than a causal mediation test.

The distributional results broaden the implications of this amplification. Recent evidence shows that educational inputs can shape cognitive endurance and socio-emotional capacities (Bruhn et al. 2025; Sorrenti et al. 2025). Our analysis neither measures these capacities directly nor identifies the causal effect of attending a higher-performing high school. Instead, it establishes a fact about how students from different prior educational environments respond to the same small shocks. Students assigned to a given examination set face the same first question and scoring rule, yet students from good high schools experience smaller performance losses. Auxiliary results suggest that this differential is specific to first-question difficulty and is not readily explained by a one-dimensional gradient in observed academic performance. When university applications are fixed before the exam, the unequal score responses also translate into differences in relative rank and STEM admission. A formally common feature of assessment design can therefore interact with prior educational advantage and redistribute opportunities among students subject to the same selection rule.

The results have broader implications for high-stakes selection. Selection systems assign explicit scoring weights to individual components, yet a question can influence final outcomes by more than its assigned score value. An unexpectedly difficult first question may differentiate students partly by how well they maintain performance following a setback, a dimension of performance that may matter beyond the examination itself. When sensitivity to the same shock varies systematically with students' prior educational environments, an incidental feature of assessment design can give this dimension disproportionate influence over selection. The relevant question is therefore not whether high-stakes examinations should contain difficult material, but whether

an unexpected opening shock should influence outcomes beyond its formal stakes and thereby alter who succeeds and how scarce educational opportunities are allocated.

The remainder of the paper proceeds as follows. Section 2 describes the institutional background of NCEE and the college application process. Section3 presents the administrative data and variable construction. Sections 4 and 5 describe empirical strategy and results. Section 6 examines robustness checks and supplementary evidence. Section 7 explores potential mechanisms, and Section 8 concludes.

## 2. Institutional Background

### 2.1. High Stakes and High Risk

The National College Entrance Examination (NCEE or *Gaokao*) is the sole gateway to higher education in China and is widely regarded as one of the highest-stakes standardized tests in the world. Admission to college—and, crucially, to *which* college—is determined almost exclusively by performance on a single examination administered once a year over two to three days.[1] The exam therefore functions as a one-shot screening device whose outcomes have been shown to carry long-run consequences for earnings and careers (Li et al. 2012; Jia and Li 2021).

The stakes are amplified by intense competition. Between 1999 and 2003, the number of exam takers rose from roughly 2.1 million to 3.7 million,[2] while admission for good colleges remained scarce. Because admission cutoffs are determined by the distribution of scores within a province–year cell, marginal score differences map into large rank displacements: around the median of the score distribution, losing a single point (out of 750) displaces a student by approximately 439 positions in the provincial ranking.

[1] Unlike the U.S. system, in which SAT or ACT scores are combined with high school grades, essays, and recommendation letters, Chinese universities during our sample period (1999–2003) admitted students essentially on the basis of the aggregate NCEE score alone, subject to students' submitted preference lists. Students who failed to gain admission could only re-take the exam a full year later.

[2] The participants number here is for science and comprehensive track. See Figure A1

### 2.2. Exam Design: Centralized Standards with Province–Year Variation

Exam content is set by the National Education Examinations Authority (NEEA) under the Ministry of Education. The NCEE is administered according to a nationally prescribed timetable. During our sample period, the examination timetable was uniform nationwide. Mathematics was administered on the afternoon of the first day for all students.[3]

Two features of exam-set assignment are central to our empirical strategy. First, the exam is *centralized and uniform within cells*: all students taking the exam in the same province, year, and academic track face an identical exam booklet.[4] Exam papers are kept strictly confidential until the exam begins,[5] so students cannot select into, anticipate, or avoid a particular exam.

Second, exam sets *vary across provinces and years*. Beginning in 2000, the NEEA moved away from a single national exam set and administered multiple exam sets in each year.[6] Because all papers are based on a standardized national syllabus, different exam sets are intended comparability; realized difficulty, however, varies across exam sets in ways that students cannot predict. This combination of centralized standards with province–year variation in exam content generates the province-year level variation in question difficulty that we exploit.

[3]Chinese was administered on the morning of the first day, the comprehensive-subject examination on the morning of the second day, and English on the afternoon of the second day. Students across provinces therefore took each subject in the same designated session.

[4]Students select into one of two curricular tracks—humanities or science—prior to the exam, and the two tracks take different mathematics exams. During our sample period the science track accounts for roughly 70 percent of exam takers; Guangdong, Guangxi and Henan operated a comprehensive track in some years. Our analysis uses science- and comprehensive-track students, for whom the mathematics curriculum and exam structure are comparable.

[5]Papers are printed and transported under national-secret protocols, and leakage is a criminal offense.

[6]The assignment of exam sets to provinces was an administrative decision driven by security considerations and the staggered rollout of curricular reform, rather than by contemporaneous local conditions. Table 1 lists the exam-set assignment for every province and year in our sample.

### 2.3. The Mathematics Exam: Design and Procedure

We focus on the mathematics exam for three reasons. First, mathematics is a core subject: it is taken by all students in all tracks, provinces, and years, and carries the maximum subject weight of 150 out of 750 aggregate points. Second, it exhibits the largest standard deviation among the three universal subjects (Chinese, Math and English), making it a key differentiator across students. Third, mathematics questions admit objective difficulty comparisons, which is essential for our measurement strategy.

Each mathematics exam in our sample follows a common architecture worth 150 points: multiple-choice questions (60 points), fill-in-the-blank questions (16 points), and multi-step problem-solving questions (74 points). [7] Our digitized corpus covers all 22 mathematics exam sets administered to science- and comprehensive-track students between 1999 and 2003, comprising 484 questions.

The first question plays a distinctive role in exam convention. By long-standing design norms—known to teachers and students alike—the opening question is intended to be the *easiest* item on the paper, serving as a warm-up that allows students to settle into the exam.[8] Scoring is binary—blank answer is scored identically to an incorrect one. Students can return to the first question and revise their answer at any time during the exam session. The point value of the first question is therefore small (at most 5 out of 150), and any effect of its difficulty on total scores beyond this mechanical component should operate through performance on the remainder of the exam.

[7] Shanghai is exception: multiple-choice questions (16 points), fill-in-the-blank questions (48 points), and multi-step problem-solving questions (86 points).

[8] No analogous convention pins down the difficulty of interior questions. In most exam sets the first question is a multiple-choice item worth 4 or 5 points; the exception is Shanghai in 2000–2003, where the exam opens with a 4-point fill-in-the-blank item.

### 2.4. Application Timing: Pre- and Post-Exam Application

Admission operates at the province level: universities allocate admission quotas to each province, and students compete only against other students in their own province–year–track cell on the basis of NCEE rank and submitted preference lists over university–major pairs.

The *timing* of preference submission varied across provinces and years in three regimes: (i) *pre-exam* application, in which students submit their complete preference list before NCEE and cannot revise it afterward; (ii) *post-exam, pre-score* application, in which students apply after the exam but before learning their scores; and (iii) *post-score* submission, in which students apply with full knowledge of their scores.[9]

This variation is central to interpreting the consequences of exam-day shocks. Under pre-exam submission, preference lists are fixed before any shock is realized: students cannot adjust their application portfolio to hedge against a lower-than-expected score, so a given score shortfall translates directly into a lower probability of clearing the cutoffs on the submitted list. Under the post-exam regimes, by contrast, students can partially adjust their targets after observing or forming beliefs about their performance. The submission regime thus governs the extent to which exam-day shocks pass through to admission outcomes.

## 3. Data, Measurement and Summary Statistics

Our analysis combines administrative records for NCEE participants with a new question-level dataset covering the mathematics examinations NCEE participants faced. The administrative records provide student outcomes but contain no responses to individual questions. We therefore construct an outcome-independent measure of question diffi-

[9]The assignment of submission regimes to province–years was an administrative choice unrelated to exam content.

culty from the content of the examination. This section describes the two data sources, including how to construct the exam difficulty dataset, and their linkage. Appendix **??** provides the complete measurement procedure and validations.

### 3.1. Administrative Data on NCEE Takers

The administrative dataset covers students approximately 13.6 million students who took the NCEE between 1999 and 2003. For each student, we observe demographics (gender, urban/rural hukou, whether is fresht graudate), exam year and province, academic track, high school identifier, raw scores by subject (Chinese, math, English, and comprehensive subjects), and college admission outcomes, including whether the student was admitted, the admitting university, and the student's major.

We restrict the analysis to students in the science and comprehensive tracks,[10] for which we observe and digitize the corresponding mathematics examinations.

### 3.2. LLM-Based Measurement and Difficulty Data

The administrative records identify the mathematics score obtained by each student but not the score of each question. Inferring difficulty from realized scores would be problematic in this setting. Score differences across provinces and years reflect not only differences in exam content but also changes in participation and in the composition of the test-taking population. Conventional item-response methods cannot resolve this problem because they require item-level responses and, for comparisons across exam forms, common anchor questions or equivalent examinee groups. None of these inputs is available for the NCEE.

We address this measurement problem by using LLM. We digitize all 484 questions from the 22 mathematics exam sets administered to the students in our sample. For each

[10] Science and comprehensive tracks account around 80% in overall sample.

question, the dataset records its text, position, type, point value, track, year, province and exam set. We then use repeated pairwise comparisons by GPT-4o to rank questions within the same track and question type, including comparisons across exam sets. For each pair, the model classifies the first question as a bit more difficult, a bit less difficult, or approximately equal in difficulty to the second. We present each pair in both orders to limit sensitivity to response order. A question receives two points for the higher difficulty level, one for a tie, and zero for the lower difficulty level; its raw difficulty index is its mean score across all eligible comparisons and ranges from zero to two. This use of LLM output as an input into empirical measurement is related to recent work using LLMs as measurement devices (Chen et al. 2024; Ludwig, Mullainathan, and Rambachan 2026). Appendix B reports the prompt, comparison design, aggregation rule, and additional implementation details.

Let $d_{ej}$ denote the resulting difficulty of question $j$ in exam set $e$, and let $s_{ej}$ denote the points assigned to that question. We define aggregate exam difficulty as the point-weighted average

$$D_e = \sum_j \left( \frac{s_{ej}}{150} \right) d_{ej}, \tag{1}$$

where $\sum_j s_{ej} = 150$. First-question difficulty is $d_{e1}$. Both measures are standardized in the regression analysis, so the estimated coefficients represent the effect of a one-standard-deviation change in the corresponding measure.

Figure 1 summarizes the resulting difficulty profiles. Two features are central to the analysis. First, opening questions are usually easy. The mean first-question difficulty is 0.447 on the raw zero-to-two scale, compared with 0.979 across all questions, and the first question is easier than the average question on its own exam set in 19 of the 22 exam sets, which is shown in Table 3. Second, first-question difficulty contains substantial variation that is distinct from the rest of the examination. Figure A3 plots first-question

difficulty against the average difficulty of the remaining questions. Its correlation with the aggregate difficulty is 0.394, and with measured difficulty of the remaining questions is 0.297. These patterns support interpreting an unusually difficult first question as a departure from the established easy-opening convention rather than merely as one component of a generally difficult exam.

We validate the LLM-based measure against four independent benchmarks. Validation results are presented in Figure 2 and Table 4. Measured difficulty rises with question position within question type, as intended by the exam design; among problem-solving questions, its correlation with normalized position is 0.952. It also rises monotonically across six difficulty categories assigned by mathematics experts, with a correlation of 0.963 between category-level expert ratings and mean LLM difficulty. At the exam level, a one-standard-deviation increase in measured aggregate difficulty predicts a 0.189-standard-deviation reduction in realized mathematics scores in the specification with province and year fixed effects and student characteristics controls. Finally, applying the same procedure to the 2026 NCEE examinations, which were not available when the model version used in our analysis was released, recovers the expected upward-sloping difficulty profile within each exam set. The full results are reported in Appendix B. We also do the robustness check for the different prompts and large language models in section 6.

The same pairwise procedure provides auxiliary measures used in the mechanism analysis. Calculation complexity proxies for the time required to solve a question. A cognitive-demand index combines the number of knowledge points,[11] their abstractness, comprehension difficulty, and innovativeness. Figure 3 plots first-question difficulty, calculation complexity, and cognitive demand across the 22 exam sets. The three

[11]A knowledge point is a distinct curricular concept, theorem, property, or standard solution technique required to solve a question. For example, a property of trigonometric functions constitutes one knowledge point, while a criterion for parallel lines in geometry constitutes another. Repeated use of the same curricular element within a question is counted only once.

measures exhibit substantial variation across exam sets and are related but not identical. We use these measures in section 7 to assess whether the additional performance loss following an adverse opening is more consistent with mechanical time depletion or cognitive interference. Since we do not observe students' time allocation or cognitive states directly, these exercises provide mechanism-consistent evidence rather than a formal mediation test. Appendix B gives the precise definitions and prompt.

We link the exam difficulty data to the administrative NCEE participants data, assigning each student the measures for the exam set she faced. Several provinces sometimes used the same exam set within a year and track. Consequently, the estimation sample contains 22 distinct realizations of the question-difficulty profile.

### 3.3. Variable Definitions

We introduce the definition of several constructed measures used in our analysis.

*NCEE Rank* is the percentile rank of a student's aggregate NCEE score relative to all test-takers in her province, year and track. The measure ranges from 0 to 1, with higher values indicating better relative performance.

*Good High School* is an indicator equal to one if a student's high school has a leave-one-cohort-out first-tier university admission rate above the 75th-percentile of the high school first-tier admission rate distribution within the province, and zero otherwise. A school's first-tier admission rate is defined as the share of its students admitted to first-tier universities. We construct this rate using all available cohorts other than the student's own cohort, so that the measure does not mechanically incorporate contemporaneous student outcomes. [12]

*College Quality* measures the academic caliber of a university's admitted students,

[12]We exclude observations for which the reported high school cannot be matched to a specific school, including students whose school is reported only as a social-candidate status or a township/locality. For schools observed in only one cohort, the leave-one-out admission rate cannot be constructed. We therefore use the contemporaneous first-tier admission rate as a fallback measure.

purged of any mechanical dependence on a given student's own cohort. We construct it in three steps. First, for each university we compute the mean NCEE rank of its admitted students, using all cohorts other than the student's own (a leave-one-cohort-out mean), so that a student's college quality measure is never mechanically correlated with her own performance. Second, within each province-year, we rank universities by this leave-one-cohort-out mean, weighting each university by the number of students it enrolls from that province-year and using midranks to break ties, so that a university enrolling more students contributes proportionally more to the ranking. Third, we convert this rank into a percentile, so that College Quality ranges from 0 to 1 and is comparable across province-years. Higher values indicate higher college quality.

*STEM College* and *First Tier College* are indicators for admission to the corresponding college types.

*Pre-Exam Submission* indicates that college applications in the student's province-year were submitted before the exam, rather than after NCEE.

*Math Difficulty* and *First Question Difficulty in Math* are the exam-set-level and first-question-level LLM-based difficulty measures described in subsection 3.2, reported on their raw 0–2 scale.

### 3.4. Summary Statistics

Table 5 reports summary statistics for our estimation sample, separately for students from good and non-good high schools, defined in subsection 3.3.

Panel A reports demographic characteristics. The sample includes around 13.6 million students, of whom 34.7% are female and 45.3% hold an urban hukou. Students in the science and comprehensive tracks account for 80.6% and 19.4% of the sample, respectively. By construction, 25.1% of students attend a good high school. Good-high-school students differ substantially from their peers on observable characteristics: they

are more likely to hold an urban hukou (64.7% versus 38.9%) and more likely to be fresh graduate test-takers (81.4% versus 75.2%), while the two groups are similarly distributed across science and comprehensive tracks.

Panel B summarizes NCEE performance. The average aggregate NCEE score is 420.8, and the average math score, our primary outcome, is 80.5 out of 150 with largest standard deviation (25.4) among three mandatory subjects. Good-high-school students substantially outperform their peers across every subject: they score 68.0 points higher on the aggregate NCEE, and 13.8 points higher on math, with correspondingly large gaps in NCEE percentile rank (66.5% versus 44.6%).

Panel C summarizes college admission outcomes. Overall, 61.2% of students are admitted to some college, 15.2% to a STEM college, and 12.5% to a first-tier college. The corresponding outcomes are substantially higher among students from good high schools. Relative to students from other high schools, they are 20.8 percentage points more likely to be admitted to college, and 21.9 percentage points more likely to be admitted to a first-tier college. Among college entrants, students from good high schools also attend higher-quality institutions, with an average College Quality of 0.625, compared with 0.446 among students from other high schools, and they are 12.7 percentage points more likely to be admitted to a STEM college.

Table 6 separately summarizes the mathematics-exam characteristics central to the analysis. The table is constructed at the province–year level and contains 142 observations. Aggregate exam difficulty has a mean of 0.983 and a standard deviation of 0.157 on the raw zero-to-two scale. First-question difficulty has a mean of 0.451, a standard deviation of 0.306, and ranges from 0 to 1.731, revealing substantial variation in how difficult the exam is at its opening. The table also reports the calculation complexity and cognitive demand of the first question, which we use in section 7 to examine potential transmission channels. Although the table is reported at the province–year level, provinces using the same exam set are assigned identical question characteristics. The

underlying variation therefore comes from 22 distinct exam sets.

Table A2 examines whether first-question difficulty is systematically related to predetermined characteristics of the test-taking population or to local economic conditions. We find no statistically significant relationship with the female share, the urban-hukou share, county GDP per capita, or the province–year admission rate. These tests provide no evidence that first-question difficulty proxies for observed student composition or contemporaneous local conditions. This evidence complements the institutional assignment of exam sets discussed in section 2.

These patterns together motivate our focus on high-school quality as the primary dimension of heterogeneity in subsection 4.2. Good- and non-good-high-school students differ substantially in educational opportunities, academic performance, and college admission outcomes, while facing similar institutional settings and realized exam difficulty. High-school quality therefore provides a natural lens through which to examine whether the effects of a negative shock at the start of the exam vary with students' prior educational advantage.

# 4. Empirical Strategy

## 4.1. Baseline Specification

To estimate the causal effect of an unexpected difficulty shock in the first question on student performance and admission outcomes, we estimate the following equation:

$$Y_{i,p,t} = \beta_0 + \beta_1 \cdot \text{Q1 Diff}_e + X_e + X_i + \gamma_t + \gamma_p + \varepsilon_p \tag{2}$$

where $i$ indexes individual students, $e$ indexes exam sets, $p$ indexes provinces, and $t$ indexes years. The outcome $Y_{i,p,t}$ includes math score, NCEE score, NCEE rank, an admission indicator, admitted university quality, and an indicator for admission to a

STEM university, all standardized to have mean zero and unit standard deviation within our sample. Q1 Diff$_e$ is our LLM-based measure of the difficulty of the first question in exam set $e$. $X_e$ is aggregate difficulty of exam set, which absorbs variation in overall exam difficulty and isolates the additional effect of the first-question shock specifically. $X_i$ is a vector of individual-level controls (gender, urban hukou status, and fresh graduates). $\gamma_t$ and $\gamma_p$ are year and province fixed effects, which absorb nationwide time trends and persistent regional differences in student composition and educational resources. Standard errors are clustered at the province level, the level at which the difficulty shock is assigned.

The coefficient of interest, $\beta_1$, captures the effect of a one-standard-deviation increase in first-question difficulty on student outcomes, holding constant the difficulty of the rest of the exam.

### 4.2. Heterogeneity Prior Educational Environment

The baseline specification estimates the average response to a difficult opening. This average may conceal important distributional consequences if students differ in their capacity to absorb the same shock. We organize the heterogeneity analysis around high school background. High schools reflect several dimensions of students' prior educational environments, including student selection, peer composition, instructional resources, and preparation for unfamiliar problems under pressure. These dimensions may shape how students respond to an unexpected setback in a high-stakes academic task.

To examine this dimension of heterogeneity, we augment the baseline specification with the Good School indicator defined in Section 3.3 and its interaction with first-question difficulty. The coefficient on first-question difficulty captures the effect among students from other high schools, while the interaction coefficient measures

the differential response among students attending good high schools.

Students within a province–year cohort face the same mathematics exam set. The interaction therefore compares responses to a common opening shock rather than differences in exposure to the shock. Since Good School status reflects both student selection and features of the school environment, we do not interpret the interaction as the causal effect of attending a good high school. Instead, it tests whether sensitivity to an adverse opening varies systematically with prior educational advantage.

### 4.3. Isolating the Performance Channel: Pre-Exam Application

The timing of university applications relative to the NCEE varies across province-years, providing a setting in which to separate the performance consequences of the exam shock from subsequent adjustments in application behavior. When applications are submitted before the exam, students make their university choices before observing their exam performance. Any effect of first-question difficulty on admission outcomes therefore operates without the post-exam margin through which students may revise their application choices in response to their realized performance.

We consequently use the pre-exam application subsample as our preferred setting for examining whether the performance shock translates into admission consequences. Within this subsample, we first re-estimate the baseline sepcification to examine the average effect of first-question difficulty. We then augment the specification with the interaction between first-question difficulty and the indicator for attending a good high school. This setting allows us to assess whether an adverse performance shock translates into differences in admission outcomes when subsequent adjustment in application choices is not available. Section 5.4 reports the results: the average performance and admission effects are presented first, followed by the heterogeneous effects by high school quality.

# 5. Empirical Results

## 5.1. Effect on Math Performance

Table 7 reports a sequence of estimates of the effect of first-question difficulty on mathematics performance. Column (1) shows an unconditional estimate of $-0.410$ standard deviations. Controlling for aggregate exam difficulty in Column (2) attenuates the estimate to $-0.264$, underscoring the importance of distinguishing an adverse opening from a aggregate difficult exam set. Adding year fixed effects in Column (3) and province fixed effects in Column (4) yields estimates of $-0.137$ and $-0.337$, respectively. With both sets of fixed effects in Column (5), the estimate is $-0.251$. Adding individual characteristics in Column (6) leaves the coefficient virtually unchanged at $-0.252$, consistent with observable student composition not driving the result.

Column (6) is our preferred specification. Its estimate of $-0.252$ standard deviations corresponds to approximately 6.399 points on a 150-point exam. Since the first question is worth only four or five points, the maximum loss it can cause directly does not account for the full estimated decline.

## 5.2. Effects on NCEE Score, Rank, and Admission Outcomes

Table 8 extends the analysis to aggregate NCEE scores, NCEE ranks, and college admission outcomes. A one-standard-deviation increase in first-question difficulty lowers the aggregate NCEE score by 0.198 standard deviations, corresponding to approximately 18.814 points out of 750. [13] Despite this decline in absolute performance, we find no detectable effects on students' relative ranks, the probability of college admission, or the likelihood of admission to a STEM university. The only evidence of an effect on

[13] Table A3 finds no evidence that first-question difficulty in mathematics spills over to performance in other subjects. These regressions are estimated at the province-year level, matching the level at which the exam set, and hence first-question difficulty, varies.

admission outcomes is a marginally significant decline in the quality of the admitting institution, with a coefficient of $-0.001$ at the 10% significance level.

This null average effect on rank and admission follows naturally from the pattern documented in Table A7: as the shock depresses scores by a similar magnitude across the distribution of math score, students' relative positions and admission outcomes are largely preserved on average even though their absolute performance is not.[14] This raises a question we turn to next: if the shock does not affect students' relative position on average, does it nonetheless affect some students more than others, potentially widening or narrowing existing gap in educational outcomes?

### 5.3. Heterogeneity by High School Quality

Table 9 examines whether students from different prior educational environments respond differently to the same opening shock. The interaction between first-question difficulty and the *Good High School* indicator is positive and statistically significant for mathematics performance. A one-standard-deviation increase in first-question difficulty reduces mathematics scores by 0.082 standard deviations less among students from good high schools than among students from other high schools. For the remaining outcomes, the interaction estimates are generally positive but not statistically significant estimated in the full sample.

As discussed in subsection 4.2, the *Good High School* indicator reflects both student selection and features of the school environment. The estimated differential could therefore arise if higher-achieving students are generally better able to absorb an adverse shock. We examine this possibility using two complementary comparisons.

[14]This near-zero average masks substantial heterogeneity in how a given score loss translates into changes in a student's position in the score distribution. In the NCEE, the number of students concentrated at a given score varies substantially across the score distribution. Around the median, a one-point difference in math score separates roughly 439 students in the same province-year-track cohort, compared with around 179 and 245 students at the 10th and 90th percentiles, respectively. Thus, the same loss in test scores can correspond to a much larger displacement in a student's position in the score distribution near the median than in the tails.

First, Table A6 divides students into the top and bottom halves of the achievement distribution within their own high-school, province–year cell. The interaction between first-question difficulty and top-half status is small and statistically insignificant. Higher-achieving students within the same school therefore do not appear systematically less sensitive to a difficult opening.

Second, Table A7 re-estimates the good-high-school differential separately across deciles of the aggregate NCEE score distribution. The interaction between first-question difficulty and the Good School indicator is positive in every decile and statistically significant in all but the bottom decile. Its magnitude is also relatively similar across most of the score distribution. The smaller performance loss among good-high-school students is therefore not confined to the highest-performing students or generated solely by their greater representation in the upper tail of the achievement distribution.

A separate exercise asks whether students from good high schools are less sensitive to difficult examinations more generally. Table A5 replaces first-question difficulty with aggregate exam difficulty. The interaction between aggregate difficulty and the *Good High School* indicator is close to zero and statistically insignificant. The relative advantage of good-high-school students is therefore specific to an adverse opening rather than a general difference in their response to difficult exam set.

These exercises do not rule out all differences in latent ability or identify which feature of the high-school environment accounts for the heterogeneity. They show, more narrowly, that the differential response is not readily explained by a monotonic relationship between observed achievement and sensitivity to difficulty. Students from different prior educational environments appear to respond differently to the same opening-specific shock. We next examine whether this differential response translates into admission consequences when application choices are fixed before the examination.

### 5.4. Pre-Exam Application Sample: Reshuffling of Admission

Table 10 restricts the sample to province-years in which applications are submitted before the exam, so that any effect on admission outcomes operates purely through realized performance. The performance effect is similar in sign and magnitude to the full sample: a one-standard-deviation increase in first-question difficulty lowers math score by 0.196 standard deviations and aggregate NCEE score by 0.091 standard deviations, and reduces the probability of admission by 1.5 percentage points and NCEE rank by a small but significant margin.

Table 11 further interacts first-question difficulty with the good-high-school indicator within the pre-exam subsample. The heterogeneous effects are more pronounced in this setting. The interaction is positive and statistically significant for both math score and aggregate NCEE score, with coefficients of 0.103 and 0.058, respectively. The differential effect also extends to NCEE rank, where the interaction coefficient is 0.029. Most notably, the interaction increases the probability of admission to a STEM university by 1.4 percentage points. Applied to the 546,273 students from good high schools in the pre-exam sample, this corresponds to roughly 7,650 additional STEM admission outcomes following a one-standard-deviation increase in first-question difficulty. This effect implies a substantial reallocation of STEM admission outcomes toward students from good high schools. With application choices fixed before the exam, these heterogeneous performance effects translate into differences in admission outcomes without subsequent application adjustment.

# 6. Robustness Checks and Supplementary Evidence

## 6.1. Restricted Samples

Our identification strategy assumes that, conditional on the standardized aggregate difficulty of the exam, and province and year fixed effects, first-question difficulty is orthogonal to unobserved student- and cohort-level shocks. Two institutional features of the NCEE could potentially threaten this assumption. First, some provinces introduce a new exam format for the first time in a given year, raising the possibility that unfamiliarity with the new format, rather than the difficulty of the opening question itself, drives our estimates. Panel A of Table A8 addresses this concern by excluding all province-years in which a new exam set was administered for the first time. The estimated effects remain negative and statistically significant for math, NCEE scores, and university quality.

Second, some province-years use province-specific exam sets rather than one of the national sets. Since our difficulty measure is constructed from the exam set administered in each province-year, these observations may raise concerns about the comparability of difficulty across province-years. Panel B therefore excludes all province-years using province-specific exam sets. These exam sets are used exclusively for science-track students, leaving a sample composed entirely of science-track test-takers. This restriction also alleviates concerns about the inclusion of comprehensive-track students in the baseline sample. The estimated effects remain negative and statistically significant for both math and NCEE scores, while the effect on NCEE rank is also negative and marginally significant. Taken together, Panels A and B provide little evidence that the estimates are driven by first-time exposure to a new exam format or the use of province-specific exam sets.

### 6.2. Alternative Prompt and LLM Model to Construct Difficulty

A potential concern is that an LLM-based difficulty measure may depend on both the prompt and the model used to construct it. We examine these two sources of sensitivity separately. First, holding the model fixed at ChatGPT-4o, we replace the baseline prompt with one that instructs the model to evaluate each question from the perspective of a typical Chinese high-school student taking the NCEE during 1999–2003 and to rely only on knowledge covered by the contemporaneous high-school curriculum. This prompt aligns the model's assessment more closely with the information set of the relevant test takers and addresses the concern that an LLM's broader mathematical knowledge may distort its evaluation of perceived difficulty. Second, holding the baseline prompt fixed, we reconstruct the difficulty measures using Claude Sonnet 4.6. These checks isolate sensitivity to prompt design and model choice, respectively.

Both alternatives exhibit substantial agreement with the baseline measures. Under the alternative prompt, the correlations with the baseline measures are 0.984 for exam difficulty in set level and 0.881 for first-question difficulty. When Claude Sonnet 4.6 is adopted, the corresponding correlations are 0.782 and 0.839, respectively. The alternative procedures therefore capture a substantial common component of question difficulty without mechanically reproducing the baseline scores.

Table A9 reports estimates obtained from the alternative measures. Panel A shows that, under the alternative prompt, a one-standard-deviation increase in first-question difficulty reduces math scores by 0.224 standard deviations and total NCEE scores by 0.199 standard deviations; both estimates are significant at the 1% level. The estimated effect on university quality is also negative and significant at the 5 percent level. Panel B replaces GPT-4o with Claude Sonnet 4.6. The estimated effect on math scores remains large and negative at $-0.306$ standard deviations and is significant at the 5% level. The Claude-based measure also yields negative effects on university admission and STEM

admission of $-0.044$ and $-0.030$, respectively. Although the remaining downstream estimates are less precise, they are uniformly nonpositive. Across both checks, the adverse effect of first-question difficulty on math performance remains economically sizable and statistically distinguishable from zero. The central finding is therefore not an artifact of either the baseline prompt or the choice of baseline LLM.

### 6.3. Alternative of Good High School

Our heterogeneity analysis classifies high schools as "good" based on their first-tier university admission rate. To assess sensitivity to this threshold, Table A10 replicates our heterogeneity specification under two alternative cutoffs: schools with a first-tier rate above the sample average (Panel A) and schools in the top decile of the first-tier rate distribution, i.e., above the 90th percentile (Panel B).

Under both definitions, the interaction between first-question difficulty and the good-high-school indicator is positive, with larger point estimates under the more selective definition. Using the average-level threshold, the interaction is statistically significant only for math scores, with a coefficient of 0.076 at the 1% level. Using the top-decile threshold, the interaction is statistically significant for math scores, NCEE scores, NCEE rank, and university quality, with coefficients of 0.112, 0.082, 0.015, and 0.016, respectively. Overall, the positive interaction is robust to alternative definitions of a good high school, suggesting that our heterogeneity results are not sensitive to the choice of thresholds.

### 6.4. Heterogeneous Effects Across Different Individual Characteristics

Beyond school quality, first-question difficulty may affect students differently depending on individual characteristics. Figure A6 presents separate estimates of our baseline specification for six mutually exclusive subgroups along three dimensions: gender,

household registration (rural vs. urban), and exam-taking status (fresh graduates or not).

The coefficient on first-question difficulty is negative for every subgroup, and its 95% confidence interval excludes zero for most of them, indicating that the qualitative finding is pervasive rather than concentrated in a single demographic. The magnitude, however, varies meaningfully across groups. The estimated effect is roughly twice as large for rural students than for urban students, even though rural students display higher baseline standardized math scores than urban students. This pattern suggests that the larger effect among rural test-takers does not simply reflect weaker academic preparation; rather, rural students appear more vulnerable to the disruptive effect of a difficult opening question than their average performance would predict.

The effect is also somewhat larger for male than for female students and for non-fresh than for fresh graduates, though both gaps are considerably more modest than the rural-urban gap. Non-fresh graduates, who are predominantly repeat test-takers, in fact score noticeably higher than fresh graduates on the standardized math test, yet this stronger average performance is not associated with a smaller estimated effect of first-question difficulty. This pattern echoes the rural-urban comparison above: groups with comparatively strong average academic performance are not systematically shielded from the disruption caused by an unexpectedly difficult opening question.

This finding contrasts with the heterogeneity documented in subsection 5.3, where students from higher-quality high schools exhibit significantly smaller performance losses from a difficult opening question. If this buffering effect primarily reflected differences in underlying academic ability, we would expect similar patterns among individual characteristics associated with stronger math performance, such as rural residence or non-fresh status. We find no such pattern, suggesting that academic ability alone does not fully explain the moderating role of school quality. Instead, the results are consistent with a role for school-level resources or exam-taking skills that help

students cope with an unexpectedly difficult opening question. Taken together, the evidence points to meaningful heterogeneity in responses to the opening-question shock that extends beyond differences in underlying academic ability.

### 6.5. Behavioral Adjustment under Alternative Application Timing

A student's ultimate admission outcome depends not only on exam score but also on the ranked list of application preferences. In some province-years, students submit their preference after the NCEE. In this setting, students can adjust their application behavior in direct response to their realized exam experience, including any disruption caused by a difficult opening question. Admission outcomes may therefore reflect both the initial performance response and subsequent adjustments in application choices. We exploit this subsample to examine whether allowing for such behavioral adjustment alters our main findings.

Table A11 restricts the sample to province-years with post-exam application and re-estimates our baseline and heterogeneity specifications. The results closely mirror our main estimates: first-question difficulty continues to lower math score and NCEE score by economically and statistically significant margins. When we further interact first-question difficulty with the good-high-school indicator, the direct effect remains similarly sized and significant, but the interaction term turns small and not significant , in contrast to the significant buffering effect documented in Section 5.4.

This pattern is consistent with the resilience-based interpretation developed above. When applications are submitted before the exam, students enter the test with their application choices already fixed, making exam performance directly consequential for their predetermined targets. A difficult opening question may therefore create stronger incentives to recover from the initial disruption, particularly for students from high-quality schools who appear better able to do so. When applications are submitted after

the exam, students retain the option to adjust their choices in light of their realized performance. This additional margin of adjustment may reduce the stakes of fully recovering from an adverse opening shock during the exam itself, potentially attenuating the relative advantage of students from high-quality schools. Consistent with this interpretation, the school-quality interaction weakens in the post-exam regime for both test performance and admission outcomes. Application flexibility may therefore alter the stakes of the opening-question shock itself, rather than merely allowing students to adjust their applications after experiencing a performance loss.

## 7. Evidence on Mechanisms

The estimated effect of first-question difficulty exceeds the maximum number of points that can be lost directly on that question. Magnitude alone, however, does not establish why the disruption extends beyond the opening item. This section organizes the mechanism evidence around two questions. First, under what conditions does opening difficulty generate an additional performance loss? Second, is that loss more closely associated with the time required to solve the opening question or with features that make it cognitively demanding and difficult to absorb?

These exercises are not a formal mediation analysis. We do not observe item-level responses, time allocation, attention, confidence, or other psychological states during the exam. The evidence instead examines whether the effect varies with question position, departures from relevant difficulty benchmarks, the extremity of the opening shock, the difficulty of the remaining exam, and the content of the opening question. We therefore interpret the results as patterns consistent with an expectation-violating opening shock, rather than as definitive identification of a single psychological mechanism.

### 7.1. The Salience of an Unexpected Opening

The first question is uniquely suited to generating an expectation-based shock for two reasons. First, by long-standing convention, the NCEE math exam begins with relatively accessible material and becomes progressively more difficult. Years of practice with mock and past exams therefore give students a common prior about the difficulty of the opening question. No comparable convention governs the second, third, or fourth questions, whose difficulty is less tightly ordered. Second, the first question is encountered before students receive any within-exam information about that year's difficulty. Its difficulty is therefore assessed against an ex-ante reference point, whereas subsequent questions are evaluated partly relative to the difficulty already experienced. The combination of a sharp prior about the opening question and the absence of any opportunity to update that prior before encountering it makes the first question particularly likely to generate a salient departure from expectations, consistent with the reference-dependent attention framework of (Bordalo, Gennaioli, and Shleifer 2012, 2022) and its implications for subsequent attention and recall (Bordalo, Gennaioli, and Shleifer 2020).

Table 12 provides a test of the salience channel. If the effect of first-question difficulty reflects a departure from an ex-ante reference point, the effect should be concentrated in the opening position. Consistent with this pattern, the coefficient on first-question difficulty is $-0.252$ and statistically significant, whereas the corresponding coefficient for Questions 2–4 is only $-0.020$ and statistically insignificant.

Figure 4 extends this analysis to the remainder of the exam. We divide the exam into four blocks—Question 1, Questions 2–4, Questions 5–8, and Questions 9–12—and estimate separate regressions using each block's average standardized difficulty. Only the first block has a statistically significant coefficient, approximately $-0.50$; the estimate for Questions 2–4 is close to zero and closely matches the Panel C result. The

estimates for Questions 5–8 and Questions 9–12 are also statistically indistinguishable from zero, although the latter block contains the exam's hardest material. The fact that the effect appears only for the opening question, rather than for other difficult parts of the exam, is consistent with the salience interpretation: the opening question is uniquely encountered against a widely shared ex-ante reference point for difficulty.

We next construct two measures of relative first-question difficulty by subtracting from first-question difficulty the average standardized difficulty of Question 1 across all exam sets and the standardized aggregate difficulty of its corresponding exam set, respectively. Both coefficients are negative and statistically significant, at $-0.225$ and $-0.501$, respectively. The larger estimate under the aggregate-exam benchmark indicates that the effect increases with the opening question's deviation from overall exam difficulty.

The heterogeneity results reinforce this pattern. The coefficient on first-question difficulty is $-0.435$ in below-median-difficulty exams and $-0.135$ in above-median-difficulty exams. A difficult opening question therefore has a substantially larger effect when the exam is otherwise easy, consistent with a reference-dependent interpretation.

The nonlinear estimates in Table 14 provide complementary evidence. Relative to an easy first question, a first question between the 50th and 90th percentiles has no statistically significant effect on total score, whereas one above the 90th percentile lowers the score by 0.513 standard deviations; the difference between the two estimates is also statistically significant. The concentration of the effect in the extreme tail is inconsistent with a smoothly increasing response to difficulty and instead suggests that sufficiently large departures from the reference point trigger the performance effect.

The evidence points to a salience-based mechanism concentrated at the opening of the exam. Across specifications, the effect depends on the opening question's deviation from the difficulty expected in the exam context, rather than on its absolute difficulty alone. This pattern is consistent with salience theory (Bordalo, Gennaioli, and Shleifer

2012, 2022), under which attributes that stand out relative to a reference point receive disproportionate weight, and with evidence that salient early information can affect subsequent attention and recall (Bordalo, Gennaioli, and Shleifer 2020). An unexpectedly difficult opening question may therefore impair subsequent performance beyond the direct cost of solving the question itself.

### 7.2. Potential Transmission Channel: Time Requirements and Cognitive Demands

The preceding evidence shows that the performance loss is concentrated among openings that sharply violate the easy-start convention. We next ask how such a salient opening may generate a loss beyond its direct score value. We consider two potential transmission channels. First, students may spend time on the opening question that would otherwise be available for the remainder of the exam. Second, the initial disruption may continue to occupy cognitive resources after students move on, interfering with subsequent problem solving. Experimental evidence shows that cognitive load can affect economic decisions and task performance (Deck and Jahedi 2015; Deck, Jahedi, and Sheremeta 2021; Drichoutis and Nayga 2020).

We do not observe students' attention, actual time allocation, or cognitive states during the exam. Table 15 therefore examines these channels indirectly using content-based characteristics of the first question. Panel A contrasts calculation complexity, our proxy for mechanical time requirements, with a composite index of cognitive demand. Calculation complexity has a small and statistically imprecise association with mathematics performance, whether entered separately or jointly with the cognitive-demand index. This provides little support for an explanation based primarily on the time mechanically required to solve the opening question, although it does not rule out changes in actual time allocation that our proxy cannot capture.

The cognitive-demand index, by contrast, is negatively and robustly associated with

performance. A one-standard-deviation increase in the index is associated with an approximately 0.12-standard-deviation decline in the mathematics score, both with and without controlling for calculation complexity. The additional loss is therefore more closely associated with openings that are cognitively difficult to absorb than with those that are merely calculation-intensive.

Panel B decomposes the index into the number of knowledge points tested, abstractness, comprehension difficulty, and innovativeness. The negative association is concentrated in abstractness, comprehension difficulty, and innovativeness, whereas the coefficient on the number of knowledge points is statistically imprecise. The pattern suggests that what matters is not simply how much material the opening question covers, but whether it presents that material in a form that is unfamiliar or difficult to interpret.

These findings sharpen the interpretation of the opening effect. The easy-start convention makes an unusually difficult first question salient. Once this initial disruption occurs, abstractness, comprehension difficulty, and innovativeness may make the question harder to process and the disruption harder to leave behind. In this interpretation, salience explains why the opening attracts disproportionate attention, while cognitive interference provides a potential channel through which its effect extends beyond the question itself. This evidence remains suggestive rather than definitive. The content measures describe characteristics of the opening question rather than realized mediators, and the administrative data do not reveal which later questions account for the additional score loss. We therefore cannot distinguish cognitive interference from related responses such as stress, reduced confidence, or persistent reallocation of attention. The results nevertheless provide little support for a purely mechanical time-cost account and are more consistent with a cognitively persistent response to an expectation-violating opening.

# 8. Conclusion

This paper shows that the formal small stakes of an opening shock need not bound its effect on final performance. Holding fixed the overall difficulty of the mathematics exam, an unusually difficult first question reduces total performance by more than the maximum number of points that can be lost directly on that question. The effect is concentrated among exceptionally difficult openings at a position governed by a strong easy-start convention, and is more closely associated with cognitively demanding than with calculation-intensive questions. These patterns are consistent with an expectation-violating opening receiving disproportionate attention and creating a disruption that is difficult to absorb. Since item-level responses, attention, and time allocation are not observed, however, the evidence neither identifies a unique psychological mechanism nor reveals precisely where within the exam the additional loss occurs.

A common shock need not have distributionally neutral consequences. Students from higher-performing high schools experience smaller performance losses, and when university applications are fixed before the exam, this differential response carries into relative rank and STEM admission. The analysis does not identify the causal effect of attending a higher-performing high school or establish that the difference reflects a single underlying trait such as resilience. It instead documents that students arriving from different educational environments absorb the same adverse opening differently. A formally common feature of an assessment can therefore change the composition of students receiving scarce educational opportunities.

The broader implication concerns how assessment design maps performance into selection. Formal scoring rules assign explicit weights to demonstrated knowledge and problem-solving ability. Yet an expectation-violating opening can place an additional, implicit weight on the capacity to maintain performance following an early setback. Our analysis does not determine whether or how strongly this capacity should be

rewarded. It shows that such a weight can arise outside the formal scoring rule and that its consequences can vary with students' prior educational environments. Evaluating an assessment therefore requires looking beyond its total difficulty and stated point allocations. A component with limited formal stakes may influence final performance by more than its assigned value, thereby changing both the attributes rewarded and the students selected. The design problem is not only how points should be assigned, but which attributes an assessment is intended to reward—and whether its structure gives those attributes the intended weights.

# References


Anaya, Lina, Nagore Iriberri, Pedro Rey-Biel, and Gema Zamarro. 2022. "Understanding Performance in Test Taking: The Role of Question Difficulty Order." *Economics of Education Review* 90: 102293.

Arold, Benjamin W., Elliott Ash, W. Bentley MacLeod, and Suresh Naidu. 2025. "Worker Rights in Collective Bargaining." Working paper.

Benedetto, Luca, Paolo Cremonesi, Andrew Caines, Paula Buttery, Andrea Cappelli, Andrea Giussani, and Roberto Turrin. 2023. "A Survey on Recent Approaches to Question Difficulty Estimation from Text." *ACM Computing Surveys* 55 (9): 1–37.

Bordalo, Pedro, Nicola Gennaioli, and Andrei Shleifer. 2012. "Salience Theory of Choice under Risk." *The Quarterly Journal of Economics* 127 (3): 1243–1285.

Bordalo, Pedro, Nicola Gennaioli, and Andrei Shleifer. 2020. "Memory, Attention, and Choice." *The Quarterly Journal of Economics* 135 (3): 1399–1442.

Bordalo, Pedro, Nicola Gennaioli, and Andrei Shleifer. 2022. "Salience." *Annual Review of Economics* 14: 521–544.

Bradley, Ralph A., and Milton E. Terry. 1952. "Rank Analysis of Incomplete Block Designs: I. The Method of Paired Comparisons." *Biometrika* 39: 324–345.

Bruhn, Jesse, Michael Gilraine, Jens Ludwig, and Sendhil Mullainathan. 2025. "Do Test Scores Misrepresent Test Results? An Item-by-Item Analysis." NBER Working Paper 34484, National Bureau of Economic Research.

Chen, Yi, Hanming Fang, Yi Zhao, and Zibo Zhao. 2024. "Recovering Overlooked Information in Categorical Variables with LLMs: An Application to Labor Market Mismatch." NBER Working Paper 32327, National Bureau of Economic Research.

Deck, Cary, and Salar Jahedi. 2015. "The Effect of Cognitive Load on Economic Decision Making: A Survey and New Experiments." *European Economic Review* 78: 97–119.

Deck, Cary, Salar Jahedi, and Roman Sheremeta. 2021. "On the Consistency of Cognitive Load." *European Economic Review* 134: 103695.

Drichoutis, Andreas C., and Jr. Nayga, Rodolfo M. 2020. "Economic Rationality under Cognitive Load." *The Economic Journal* 130 (632): 2382–2409.

Duquennois, Claire. 2022. "Fictional Money, Real Costs: Impacts of Financial Salience on Disadvantaged Students." *American Economic Review* 112 (3): 798–826.

Ebenstein, Avraham, Victor Lavy, and Sefi Roth. 2016. "The Long-Run Economic Consequences of High-Stakes Examinations: Evidence from Transitory Variation in Pollution." *American Economic Journal: Applied Economics* 8 (4): 36–65.

Embretson, Susan E., and Steven P. Reise. 2000. *Item Response Theory for Psychologists*.: Lawrence Erlbaum.

Graff Zivin, Joshua, Yingquan Song, Qu Tang, and Peng Zhang. 2020. "Temperature and High-Stakes Cognitive Performance: Evidence from the National College Entrance Examination in China." *Journal of Environmental Economics and Management* 104: 102365.

Hambleton, Ronald K., and Hariharan Swaminathan. 1985. *Item Response Theory: Principles and Applications*.: Kluwer.

Heissel, Jennifer A., Emma K. Adam, Jennifer L. Doleac, David N. Figlio, and Jonathan Meer. 2021. "Testing, Stress, and Performance: How Students Respond Physiologically to High-Stakes Testing." *Education Finance and Policy* 16 (2): 183–208.

Holland, Paul W., and Neil J. Dorans. 2006. "Linking and Equating." In *Educational Measurement*, edited by Robert L. Brennan 4th ed.: Praeger.

Jia, Ruixue, and Hongbin Li. 2021. "Just Above the Exam Cutoff Score: Elite College Admission and Wages in China." *Journal of Public Economics* 196: 104371.

Kolen, Michael J., and Robert L. Brennan. 2014. *Test Equating, Scaling, and Linking: Methods and Practices*. 3rd ed.: Springer.

Li, Hongbin, Lingsheng Meng, Xinzheng Shi, and Binzhen Wu. 2012. "Does Attending Elite Colleges Pay in China?" *Journal of Comparative Economics* 40 (1): 78–88.

van der Linden, Wim J., and Ronald K. Hambleton. edited 1997. *Handbook of Modern Item Response Theory*.: Springer.

Lord, Frederic M. 1980. *Applications of Item Response Theory to Practical Testing Problems*.: Routledge.

Lord, Frederic M., and Melvin R. Novick. 1968. *Statistical Theories of Mental Test Scores*.: Addison-Wesley.

Ludwig, Jens, Sendhil Mullainathan, and Ashesh Rambachan. 2026. "Large Language Models: An Applied Econometric Framework." *Annual Review of Economics* 18: 283–316.

OpenAI. 2023. "GPT-4 Technical Report."Technical report, OpenAI.

Rasch, Georg. 1960. *Probabilistic Models for Some Intelligence and Attainment Tests*. Copenhagen: Danish Institute for Educational Research.

Reyes, Germán. 2023. "Cognitive endurance, talent selection, and the labor market returns to human capital.." *arXiv preprint arXiv:2301.02575*.

Rodrik, Dani, and Mert Geyiktepe. 2026. "The China Backlash: Quantifying the Narrative Discourse on China." CEPR Discussion Paper DP21787, CEPR Press.

Sievertsen, Hans Henrik, Francesca Gino, and Marco Piovesan. 2016. "Cognitive Fatigue Influences Students' Performance on Standardized Tests." *Proceedings of the National Academy of*

*Sciences* 113 (10): 2621–2624.

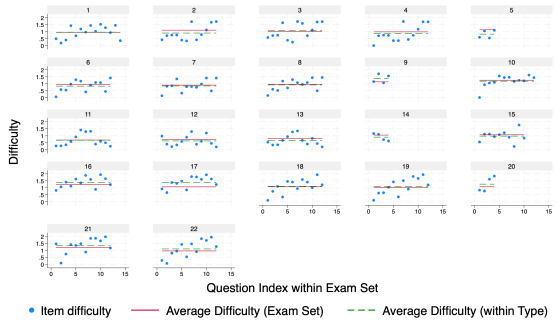
Sorrenti, Giuseppe, Ulf Zölitz, Denis Ribeaud, and Manuel Eisner. 2025. "The Causal Impact of Socio-Emotional Skills Training on Educational Success." *The Review of Economic Studies* 92 (1): 506–552.

Yuan, Zheng, Hongyi Yuan, Chuanqi Tan, Wei Wang, and Songfang Huang. 2023. "How Well Do Large Language Models Perform in Arithmetic Tasks?" *arXiv*.

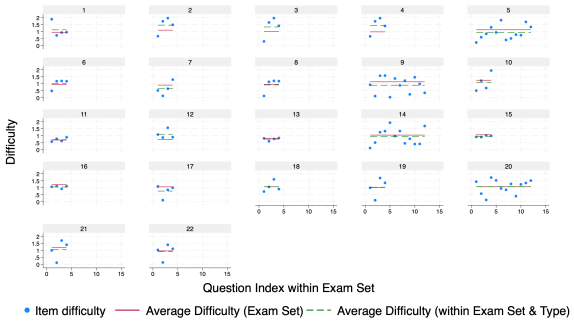
Zhang, Xiaotian, Chunyang Li, Yi Zong, Zhengyu Ying, Liang He, and Xipeng Qiu. 2023. "Evaluating the Performance of Large Language Models on Gaokao Benchmark." *arXiv*.

Zheng, Lianmin, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin et al. 2023. "Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena." In *Advances in Neural Information Processing Systems*, vol. 36: 46595–46623.

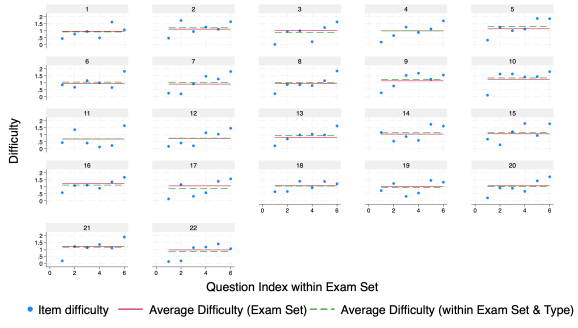
FIGURE 1. Item and Exam Set Difficulty



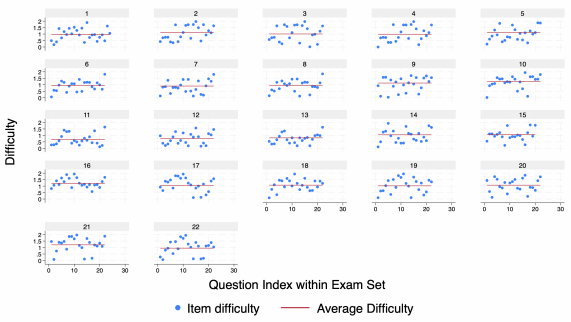

A. Multiple-Choice Questions




B. Fill-in-Blank Questions



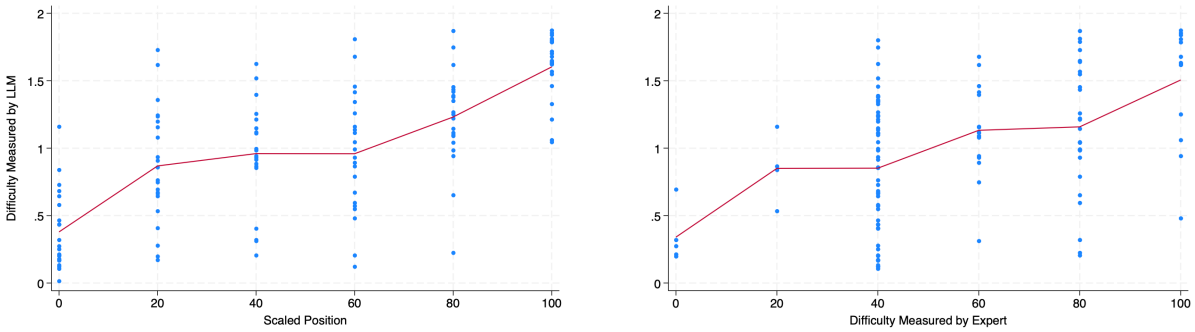

C. Problem-Solving Questions




D. All Types

*Notes:* This figure plots item difficulty by question index within each exam set, separately for multiple-choice (Panel A), fill-in-the-blank (Panel B), problem-solving questions (Panel C) and all types of questions together (Panel D). Each small panel within a subfigure corresponds to one exam set, and each dot represents a single question, ordered by its index within that exam set. Difficulty is measured in raw, unstandardized units, with higher values indicating harder items. The solid line in each panel denotes the average difficulty within that exam set, while the dashed line denotes the average difficulty within that exam set and question type. Exam sets contain between 20 and 25 questions. Across most exam sets and types, the first item exhibits below-average difficulty, consistent with the practice of opening an exam with easier questions.

FIGURE 2. Validation of the LLM-Measured Difficulty

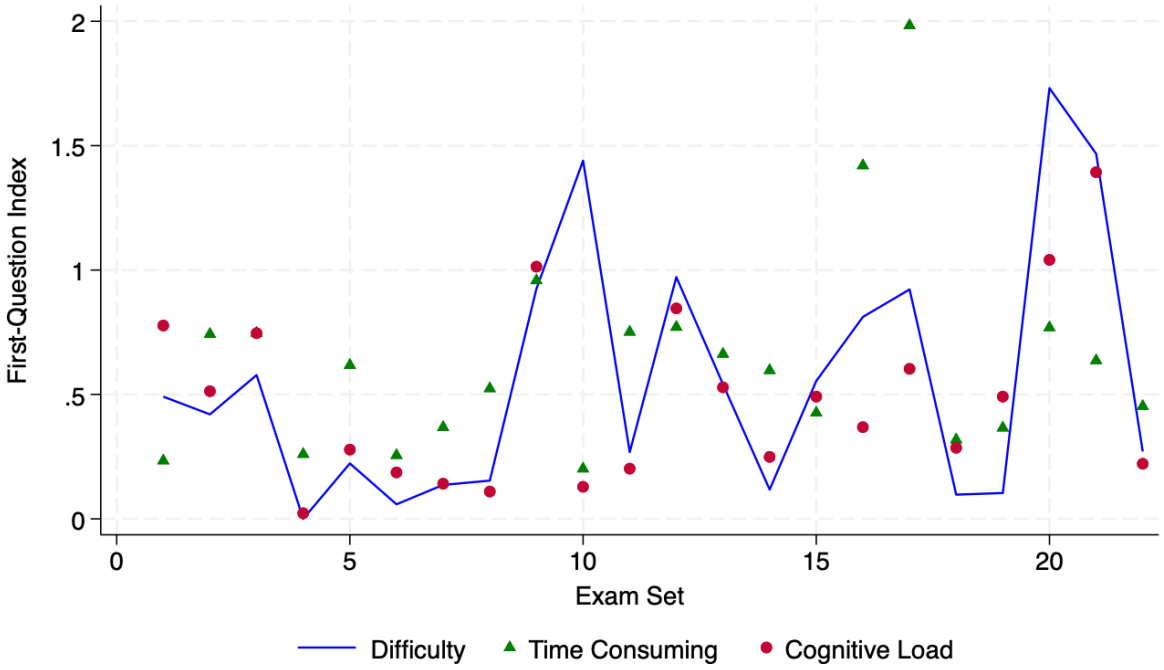



A. Question Sequence of Problem-Solving

B. Expert Assessment of Problem-Solving

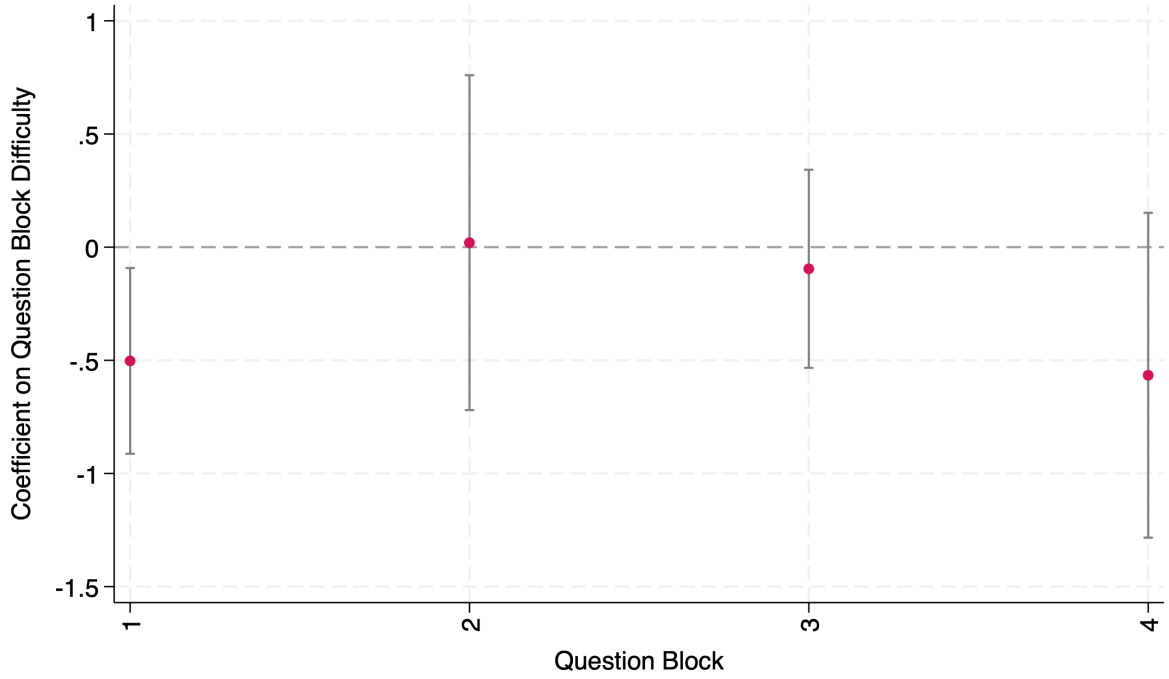
*Notes:* This figure validates the LLM-measured item difficulty, pooling problem-solving questions from all exam sets. Difficulty is in raw, unstandardized units, with higher values indicating harder items. Each dot is a single question, and the red lines connect group averages. Panel A plots difficulty against question position, rescaled so that the first question equals 0 and the last equals 100; the correlation between difficulty and position is 0.669 at the item level and 0.952 across position averages. Panel B presents difficulty against expert assessments, which classify each question into six categories mapped onto a 0–100 scale, from easy (0) to difficult (100); the correlation is 0.489 at the item level and 0.963 across expert-rating averages.

FIGURE 3. First-Question Index Across Exam Sets




*Notes:* This figure presents three indexes of first question for each of NCEE math exam sets in our sample: difficulty (solid line), time consumption (triangles), and cognitive load (circles). Each index ranges from 0 to 2; higher values indicate a more demanding opening question. These indexes are positively correlated but far from identical.

FIGURE 4. Salience Evidence: Only the First Question Difficulty Significantly Reduces Math Scores




*Notes*: This figure exhibits the estimated effects of question-block difficulty on the standardized math score. The exam questions are divided into four blocks: Question 1, Questions 2–4, Questions 5–8, and Questions 9–12. These questions are multiple-choice in most provinces, except in Shanghai, where they are fill-in-the-blank questions. Block difficulty is the raw average LLM-measured difficulty of the questions in each block. Each dot represents the coefficient on the block difficulty from a separate regression, controlling for the aggregate difficulty, individual characteristics, and province and year fixed effects. Vertical lines indicate 95% confidence intervals, with standard errors clustered at the province level.

TABLE 1. Exam Sets Across Provinces and Years

| Province | 1999 | 2000 | 2001 | 2002 | 2003 | Province | 1999 | 2000 | 2001 | 2002 | 2003 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Beijing | 1 | 2 | 6 | 10 | 15 | Hubei | 1 | 2 | 6 | 11 | 16 |
| Tianjin | 1 | 3 | 7 | 12 | 17 | Hunan | 1 | 2 | 6 | 11 | 16 |
| Hebei | 1 | 2 | 6 | 11 | 16 | Guangdong | 1 | 4 | 8 | 13 | 18 |
| Shanxi | 1 | 3 | 7 | 12 | 17 | Guangxi | 1 | 2 | 6 | 13 | 18 |
| Inner Mengolia | 1 | 2 | 6 | 11 | 16 | Hainan | 1 | 2 | 6 | 11 | 16 |
| Liaoning | 1 | 2 | 6 | 11 | 19 | Chongqing | 1 | 2 | 6 | 11 | 16 |
| Jilin | 1 | 2 | 6 | 11 | 16 | Sichuan | 1 | 2 | 6 | 11 | 16 |
| Heilongjiang | 1 | 2 | 6 | 11 | 17 | Guizhou | 1 | 2 | 6 | 11 | 16 |
| Shanghai | 1 | 5 | 9 | 14 | 20 | Yunnan | 1 | 2 | 6 | 11 | 16 |
| Jiangsu | 1 | 2 | 6 | 11 | 21 | Tibet | 1 | 2 | 6 | 11 | 16 |
| Zhejiang | 1 | 2 | 6 | 11 | 16 | Shaanxi | 1 | 2 | 6 | 11 | 16 |
| Anhui | 1 | 2 | 6 | 11 | 17 | Gansu | 1 | 2 | 6 | 11 | 16 |
| Fujian | 1 | 2 | 6 | 11 | 16 | Qinghai | 1 | 2 | 6 | 11 | 17 |
| Jiangxi | 1 | 3 | 7 | 12 | 17 | Ningxia | 1 | 2 | 6 | 11 | 16 |
| Shandong | 1 | 2 | 6 | 11 | 17 | Xinjiang | 1 | 2 | 6 | 11 | 16 |
| Henan | 1 | 2 | 8 | 13 | 22 | | | | | | |

*Notes:* This table reports the exam set in each province–year cell. Each number identifies a distinct exam set, from 1 to 22; cells sharing the same number (and color) indicate that those provinces used the identical exam set in that year. The table covers the science and comprehensive track. Information on exam set assignment is collected by the authors from the website of the Ministry of Education and websites dedicated to the NCEE.

TABLE 2. Submission Timing Across Provinces and Years

| Province | 1999 | 2000 | 2001 | 2002 | 2003 | Province | 1999 | 2000 | 2001 | 2002 | 2003 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Beijing | 2 | 3 | 1 | 1 | 3 | Hubei | 2 | 2 | 2 | 2 | 2 |
| Tianjin | 2 | 2 | 2 | 2 | 2 | Hunan | 2 | 2 | 3 | 3 | 3 |
| Hebei | 3 | 3 | 3 | 3 | 3 | Guangdong | 3 | 1 | 1 | 1 | 1 |
| Shanxi | 2 | 2 | 2 | 2 | 2 | Guangxi | 2 | 2 | 2 | 2 | 2 |
| Inner Mengolia | 2 | 2 | 2 | 2 | 3 | Hainan | 2 | 2 | 2 | 2 | 3 |
| Liaoning | 2 | 2 | 2 | 2 | 2 | Chongqing | 2 | 1 | 1 | 1 | 1 |
| Jilin | 2 | 3 | 1 | 1 | 2 | Sichuan | 2 | 2 | 2 | 1 | 2 |
| Heilongjiang | 1 | 2 | 2 | 2 | 2 | Guizhou | 1 | 1 | 1 | 1 | 2 |
| Shanghai | 1 | 1 | 1 | 1 | 1 | Yunnan | NA | NA | 2 | 2 | NA |
| Jiangsu | 2 | 1 | 1 | 2 | 3 | Tibet | NA | NA | 1 | 1 | 3 |
| Zhejiang | 3 | 3 | 3 | 3 | 3 | Shaanxi | 2 | 2 | 2 | 2 | 2 |
| Anhui | 2 | 2 | 2 | 2 | 2 | Gansu | NA | NA | NA | NA | 2 |
| Fujian | 2 | 2 | 2 | 2 | 2 | Qinghai | 3 | 3 | 3 | 3 | 3 |
| Jiangxi | 2 | 2 | 2 | 2 | 2 | Ningxia | NA | NA | 3 | 3 | 3 |
| Shandong | 2 | 2 | 2 | 3 | 2 | Xinjiang | 2 | 2 | 2 | 2 | 2 |
| Henan | 2 | 2 | 2 | 2 | 2 | | | | | | |

*Notes:* This table presents the timing of NCEE application in each province–year cell. A value of 1 indicates that applications were submitted before the exam; 2 indicates submission after the exam but before scores were released; 3 indicates submission after scores were known. NA denotes province–year cells for which the submission timing could not be determined. Information on submission timing is collected by the authors from the website of the Ministry of Education and websites dedicated to NCEE, supplemented by direct email correspondence with provincial education authorities.

TABLE 3. Measurement Result: Aggregate Difficulty & First Question Difficulty

| Agg. Rank | Q1 Rank | Exam Set | Year | Province | Agg. Diff. | Q1 Diff. | ΔDifficulty |
|---|---|---|---|---|---|---|---|
| 1 | 3 | 10 | 2002 | Beijing | 1.2450 | 1.4394 | 0.1944 |
| 2 | 13 | 16 | 2003 | Old standard | 1.2192 | 0.2706 | -0.9486 |
| 3 | 4 | 21 | 2003 | Jiangsu | 1.2098 | 0.9719 | -0.2379 |
| 4 | 5 | 9 | 2001 | Shanghai | 1.1327 | 0.9244 | -0.2083 |
| 5 | 16 | 5 | 2000 | Shanghai | 1.1313 | 0.1537 | -0.9776 |
| 6 | 12 | 2 | 2000 | Old standard | 1.1085 | 0.4199 | -0.6886 |
| 7 | 22 | 18 | 2003 | Guangdong & Guangxi | 1.0752 | 0.0000 | -1.0752 |
| 8 | 1 | 20 | 2003 | Shanghai | 1.0732 | 0.0974 | -0.9758 |
| 9 | 11 | 15 | 2003 | Beijing | 1.0670 | 0.4913 | -0.5757 |
| 10 | 6 | 17 | 2003 | New standard | 1.0658 | 0.9221 | -0.1437 |
| 11 | 17 | 14 | 2002 | Shanghai | 1.0478 | 0.1364 | -0.9114 |
| 12 | 18 | 19 | 2003 | Liaoning | 1.0201 | 0.1176 | -0.9025 |
| 13 | 8 | 3 | 2000 | New standard | 1.0086 | 0.5779 | -0.4307 |
| 14 | 19 | 4 | 2000 | Guangdong | 0.9837 | 0.1039 | -0.8798 |
| 15 | 20 | 8 | 2001 | Guangdong & Henan | 0.9544 | 0.0974 | -0.8570 |
| 16 | 14 | 22 | 2003 | Henan | 0.9538 | 0.2684 | -0.6854 |
| 17 | 21 | 6 | 2001 | Old standard | 0.9521 | 0.0584 | -0.8937 |
| 18 | 10 | 1 | 1999 | Old standard | 0.9501 | 1.7311 | 0.7810 |
| 19 | 15 | 7 | 2001 | New standard | 0.8998 | 0.2227 | -0.6771 |
| 20 | 7 | 13 | 2002 | Guangdong & Henan & Guangxi | 0.8124 | 0.8117 | -0.0007 |
| 21 | 2 | 12 | 2002 | New standard | 0.7433 | 1.4675 | 0.7242 |
| 22 | 9 | 11 | 2002 | Old standard | 0.6932 | 0.5541 | -0.1391 |

*Notes: Agg. Rank* and *Q1 Rank* rank the 22 exam sets by difficulty, from 1 (most difficult) to 22 (least difficult), according to aggregate difficulty and first-question difficulty, respectively. *Agg. Diff.* is the weighted average difficulty across all questions in an exam set, and *Q1 Diff.* is the difficulty of its opening question; both are raw (non-standardized) difficulty scores. $\Delta Difficulty = Q1\ Diff. - Agg.\ Diff.$, so that positive values (highlighted in orange) indicate that the first question is more difficult than the exam set as a whole. *Old standard* and *New standard* denote joint exam sets administered simultaneously in multiple provinces under the old and new national curriculum standards, respectively.

TABLE 4. Math Score versus LLM Measured Difficulty

| | **Math Score** | | | | |
|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) |
| Aggregate Difficulty | -0.204*** | -0.086 | -0.224*** | -0.170*** | -0.189*** |
| | (0.039) | (0.099) | (0.017) | (0.037) | (0.040) |
| Mean of DV | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| Observations | 142 | 142 | 142 | 142 | 142 |
| R-squared | 0.161 | 0.280 | 0.746 | 0.853 | 0.859 |
| Year FE | | ✓ | | ✓ | ✓ |
| Province FE | | | ✓ | ✓ | ✓ |
| Demographic Control | | | | | ✓ |

*Notes:* This table validates the LLM-measured standardized difficulty by relating it to realized exam performance. Each observation is a province–year cell. All variables are standardized. Standard errors clustered at the province level are reported in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

## TABLE 5. Summary Statistics

| Variable | Overall | | | Good High School | | | Non-Good High School | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Observations | Mean | Std. Dev. | Observations | Mean | Std. Dev. | Observations | Mean | Std. Dev. | Difference |
| | **A. Demographic Information** | | | | | | | | | |
| Female | 13,648,730 | .347 | .476 | 3,140,185 | .368 | .482 | 9.383,757 | .341 | .474 | .027*** |
| Urban | 13,648,584 | .453 | .498 | 3,140,184 | .647 | .478 | 9.383,634 | .389 | .488 | .257*** |
| Fresh Graduate | 13,648,730 | .739 | .439 | 3,140,185 | .814 | .389 | 9.383,757 | .752 | .432 | .062*** |
| Science Track | 13,648,912 | .806 | .395 | 3,140,185 | .877 | .328 | 9.383,757 | .786 | .410 | .091 |
| Comprehensive Track | 13,648,912 | .194 | .395 | 3,140,185 | .123 | .328 | 9.383,757 | .214 | .410 | -.091 |
| Good High School | 12,523,942 | .251 | .433 | 3,140,185 | | | 9,289,453 | | | |
| | **B. NCEE Performance** | | | | | | | | | |
| NCEE Score | 13,648,488 | 420.848 | 95.018 | 3,140,110 | 471.662 | 87.118 | 9,383,614 | 403.705 | 91.897 | 67.957*** |
| NCEE Rank | 13,648,730 | .500 | .289 | 3,140,185 | .665 | .261 | 9.383,757 | .446 | .278 | .219*** |
| Chinese Score | 13,648,500 | 92.566 | 15.501 | 3,140,120 | 98.964 | 14.129 | 9.383,607 | 90.512 | 15.391 | 8.452*** |
| Math Score | 13,648,449 | 80.549 | 25.392 | 3,140,112 | 90.760 | 23.994 | 9.383,580 | 76.915 | 24.981 | 13.844*** |
| English Score | 13,647,412 | 80.138 | 22.722 | 3,140,006 | 92.060 | 20.997 | 9.383,732 | 76.012 | 21.845 | 16.048*** |
| Comprehensive Subjects Score | 13,634,613 | 167.502 | 52.235 | 3,135,225 | 189.694 | 49.188 | 9,375,561 | 160.202 | 51.009 | 29.492*** |
| | **C. College Admission Dummies** | | | | | | | | | |
| Admitted | 13,648,657 | .612 | .487 | 3,140,185 | .765 | .424 | 9.383,757 | .557 | .497 | .208*** |
| College Quality | 8,347,803 | .500 | .289 | 2,403,121 | .626 | .270 | 5,231,196 | .446 | .280 | .181*** |
| STEM College | 13,648,912 | .152 | .359 | 3,140,185 | .247 | .431 | 9.383,757 | .120 | .325 | .127*** |
| First Tier College | 13,648,657 | .125 | .331 | 3,140,185 | .290 | .454 | 9.383,757 | .072 | .258 | .219*** |

*Notes:* The sample is restricted to NCEE participants who chose science and comprehensive tracks. Panel A presents demographic characteristics. *Urban* indicates students holding an urban *hukou*; *Fresh Graduate* indicates students taking the NCEE in the year of high school graduation rather than retaking it; Panel B reports NCEE performance. All scores are reported in raw points: the total *NCEE Score* is out of 750, the *Chinese Score*, *Math Score*, and *English Score* are each out of 150, and the *Comprehensive Subjects Score* is out of 300. *NCEE Rank* is the percentile rank of the NCEE score within the province–year, ranging from 0 to 1. Panel C summarizes college admission outcomes. *Admitted* indicates admission by any college; *College Quality* is the leave-one-out mean of province–year NCEE rankings of colleges; *STEM College* and *First Tier College* indicate admission to the corresponding college types. The smaller number of observations for *College Quality* reflects that this variable is defined only for students admitted to college. Panel D reports institutional features of the NCEE faced by each student. *Pre-Exam Submission* indicates that college applications in the student's province–year were submitted before the exam; The last column reports differences in means between good and non-good high schools. Standard errors for the differences are clustered at the province level. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

TABLE 6. Summary Statistics on Mathematics Exam Design

| Variable | Overall (Year & Province Level) | | | | |
|---|---|---|---|---|---|
| | Observations | Mean | Std. Dev. | Min | Max |
| Aggregate Difficulty (Set Level) | 142 | .983 | .157 | .693 | 1.250 |
| First-Question Difficulty | 142 | .451 | .306 | .000 | 1.731 |
| First-Question Time Consumption | 142 | .663 | .462 | .201 | 1.983 |
| First-Question Cognitive Load | 142 | .466 | .258 | .023 | 1.393 |

*Notes:* This table exhibits summary statistics for mathematics-exam design across 142 province-year observations during 1999–2003. *Aggregate Difficulty* is the weighted mean LLM-assessed difficulty of all questions in an exam set, *First-Question Difficulty* is the difficulty of the opening question. *First-Question Time Consumption* is proxied by the calculation complexity of the opening question. *First-Question Cognitive Load* is a composite measure based on four dimensions: the number of knowledge points required, the degree of novelty, the abstractness of the knowledge involved, and the difficulty of understanding the question. All measures are constructed using the baseline ChatGPT-4o procedure and reported on their original scale from 0 to 2, with higher values indicating greater difficulty, calculation complexity, or cognitive load. The unit of observation is the province-year, the level at which exam-set assignment varies. Provinces administering the same exam in a given year are therefore assigned identical exam characteristics.

TABLE 7. A More Difficult First Question Lowers Math Scores

| | **Math Score** | | | | | |
|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) |
| First Question Difficulty | -0.410*** | -0.264*** | -0.137* | -0.337*** | -0.251** | -0.252** |
| | (0.000) | (0.000) | (0.077) | (0.041) | (0.103) | (0.101) |
| Mean of DV | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| Observations | 13,648,449 | 13,648,449 | 13,648,499 | 13,648,499 | 13,648,499 | 13,648,306 |
| R-squared | 0.069 | 0.103 | 0.128 | 0.212 | 0.229 | 0.241 |
| Aggregate Difficulty Control | | ✓ | ✓ | ✓ | ✓ | ✓ |
| Year FE | | | ✓ | | ✓ | ✓ |
| Province FE | | | | ✓ | ✓ | ✓ |
| Individual Control | | | | | | ✓ |

*Notes:* This table reports the effect of standardized first question difficulty on individual standardized math score. Each observation is an individual exam taker. All variables are standardized. Standard errors clustered at the province level are reported in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

TABLE 8. The Effect of First Question Difficulty on NCEE Outcome and Admission Result

| | (1) | (2) | (3) | (4) | (5) |
|---|---|---|---|---|---|
| | NCEE Score | NCEE Rank | Admitted | Uni Quality | STEM Uni |
| First Question Difficulty | -0.198*** | -0.000 | -0.030 | -0.001* | -0.012 |
| | (0.067) | (0.000) | (0.020) | (0.000) | (0.010) |
| Mean of DV | 0.000 | 0.500 | 0.612 | 0.500 | 0.248 |
| Observations | 13,648,342 | 13,648,584 | 13,643,511 | 8,347,703 | 8,347,703 |
| R-squared | 0.129 | 0.015 | 0.084 | 0.010 | 0.043 |
| Aggregate Difficulty Control | ✓ | ✓ | ✓ | ✓ | ✓ |
| Year FE | ✓ | ✓ | ✓ | ✓ | ✓ |
| Province FE | ✓ | ✓ | ✓ | ✓ | ✓ |
| Individual Control | ✓ | ✓ | ✓ | ✓ | ✓ |

*Notes:* This table presents the effect of standardized first question difficulty on NCEE outcomes beyond the math exam. The dependent variables are the standardized NCEE score in Column (1), the individual's percentile rank of the NCEE score (calculated within province–year–track) in Column (2), an indicator for admission into any college in Column (3), the quality rank of the admitted college in Column (4), and an indicator for admission into a STEM university in Column (5); Columns (4)–(5) are restricted to admitted students. All columns control for the aggregate difficulty of the math exam set, year and province fixed effects, and individual controls. Standard errors clustered at the province level are reported in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

TABLE 9. Heterogeneous Effects by High School Quality

| | (1) | (2) | (3) | (4) | (5) | (6) |
|---|---|---|---|---|---|---|
| | Math Score | NCEE Score | NCEE Rank | Admitted | Uni Quality | STEM Uni |
| First Question Difficulty | -0.299*** | -0.228*** | -0.007* | -0.034 | -0.009** | -0.010 |
| | (0.100) | (0.068) | (0.004) | (0.022) | (0.004) | (0.011) |
| Good High School | 0.630*** | 0.785*** | 0.249*** | 0.182*** | 0.200*** | 0.089*** |
| | (0.028) | (0.031) | (0.011) | (0.009) | (0.009) | (0.006) |
| First Question Difficulty × Good High School | 0.082*** | 0.047 | 0.012 | -0.006 | 0.014 | 0.004 |
| | (0.028) | (0.029) | (0.009) | (0.011) | (0.009) | (0.004) |
| Mean of DV | -0.006 | -0.001 | 0.501 | 0.610 | 0.503 | 0.249 |
| Observations | 12,523,569 | 12,523,600 | 12,523,818 | 12,523,818 | 7,634,228 | 7,634,228 |
| R-squared | 0.303 | 0.230 | 0.136 | 0.109 | 0.097 | 0.051 |
| Aggregate Difficulty Control | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Year FE | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Province FE | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Individual Control | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |

*Notes:* This table examines whether the effect of standardized first question difficulty of the math exam differs by high school quality. Both math score and NCEE score are standardized. The sample is restricted to individuals with non-missing high school information. All columns control for the aggregate difficulty of the math exam set, year and province fixed effects, and individual controls. Standard errors clustered at the province level are reported in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

TABLE 10. Pre-Exam Application: Negative Shock on Score

| | (1) | (2) | (3) | (4) | (5) | (6) |
|---|---|---|---|---|---|---|
| | Math Score | NCEE Score | NCEE Rank | Admitted | Uni Quality | STEM Uni |
| First Question Difficulty | -0.196*** | -0.091*** | -0.001* | -0.015* | -0.000 | -0.005 |
| | (0.029) | (0.016) | (0.001) | (0.008) | (0.000) | (0.003) |
| Mean of DV | 0.122 | -0.039 | 0.501 | 0.712 | 0.500 | 0.219 |
| Observations | 2,208,492 | 2,208,474 | 2,208,492 | 2,208,492 | 1,573,318 | 1,573,318 |
| R-squared | 0.156 | 0.064 | 0.008 | 0.046 | 0.008 | 0.058 |
| Aggregate Difficulty Control | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Year FE | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Province FE | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Individual Control | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |

*Notes:* This table reports the effect of standardized first question difficulty of the math exam on NCEE outcomes for the subsample of provinces–years in which college applications were submitted *before* the exam. Both math score and NCEE score are standardized. In this regime, application choices cannot respond to exam performance, so any effect on admission outcomes operates through scores alone. All columns control for the aggregate difficulty of the math exam set, year and province fixed effects, and individual controls. Standard errors clustered at the province level are reported in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

TABLE 11. Pre-Exam Application: Shuffle Composition at STEM University

| | (1) | (2) | (3) | (4) | (5) | (6) |
|---|---|---|---|---|---|---|
| | Math Score | NCEE Score | NCEE Rank | Admitted | Uni Quality | STEM Uni |
| First Question Difficulty | -0.262*** | -0.119*** | -0.012** | -0.029** | -0.006 | -0.007 |
| | (0.031) | (0.019) | (0.004) | (0.011) | (0.004) | (0.005) |
| Good High School | 0.750*** | 0.905*** | 0.289*** | 0.183*** | 0.233*** | 0.094*** |
| | (0.035) | (0.031) | (0.013) | (0.018) | (0.012) | (0.007) |
| First Question Difficulty × Good High School | 0.103*** | 0.058* | 0.019** | 0.006 | 0.011 | 0.014* |
| | (0.012) | (0.025) | (0.008) | (0.010) | (0.009) | (0.007) |
| Mean of DV | 0.122 | -0.024 | 0.502 | 0.714 | 0.503 | 0.215 |
| Observations | 2,015,767 | 2,015,761 | 2,015,767 | 2,015,767 | 1,439,047 | 1,439,047 |
| R-squared | 0.255 | 0.209 | 0.166 | 0.077 | 0.119 | 0.067 |
| Aggregate Difficulty Control | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Year FE | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Province FE | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Individual Control | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |

*Notes:* This table examines whether the effect of standardized first question difficulty of the math exam differs by high school quality, for the subsample of provinces–years in which college applications were submitted *before* the exam. Both math score and NCEE score are standardized. All columns control for the aggregate difficulty of the math exam set, year and province fixed effects, and individual controls. Standard errors clustered at the province level are reported in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

TABLE 12. Salience Channel

| | Math Score | | | |
|---|---|---|---|---|
| | (1) | (2) | (3) | (4) |
| Q1 Difficulty | -0.252** | | | |
| | (0.101) | | | |
| Q2-Q4 Average Difficulty | | -0.020 | | |
| | | (0.362) | | |
| Q1 Diff. Deviation from Q1 Ave. Diff. | | | -0.225** | |
| | | | (0.090) | |
| Q1 Diff. Deviation from Agg. Diff. | | | | -0.501** |
| | | | | (0.187) |
| R-squared | 0.241 | 0.234 | 0.241 | 0.234 |
| Mean of DV | 0.000 | 0.000 | 0.000 | 0.000 |
| Observations | 13,648,306 | 13,648,306 | 13,648,306 | 13,648,306 |
| Aggregate Difficulty Control | ✓ | ✓ | ✓ | ✓ |
| Year FE | ✓ | ✓ | ✓ | ✓ |
| Province FE | ✓ | ✓ | ✓ | ✓ |
| Individual Control | ✓ | ✓ | ✓ | ✓ |

Notes: This table tests the salience channel: Column (1) uses the standardized difficulty of the first question; Column (2) uses the standardized average difficulty of Questions 2–4; Columns (3) and (4) use the deviation of first-question difficulty from the average difficulty of the first questions and from the aggregate difficulty of the exam set, respectively. All variables in this table are standardized. All columns control for the aggregate difficulty of the math exam set, year and province fixed effects, and individual controls. Standard errors clustered at the province level are reported in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

TABLE 13. Salience Channel: Easy versus Hard Exam Set

| | **Math Score** | |
|---|---|---|
| | (1) | (2) |
| | Easier Exam ( Agg. Diff.$\leq$ 50*th* percentile) | Harder Exam ( > 50*th* percentile) |
| First Question Difficulty | -0.435*** | -0.135*** |
| | (0.053) | (0.045) |
| Mean of DV | 0.176 | -0.278 |
| Observations | 8,351,739 | 5,296,567 |
| R-squared | 0.199 | 0.251 |
| Aggregate Difficulty Control | ✓ | ✓ |
| Year FE | ✓ | ✓ |
| Province FE | ✓ | ✓ |
| Individual Control | ✓ | ✓ |

*Notes*: This table explores the salience channel by splitting the sample based on the aggregate difficulty of the math exam. All variables are standardized. Column (1) restricts the sample to province–years in which the aggregate difficulty of the math exam is below the sample median (easier exams); Column (2) restricts the sample to province–years in which the aggregate difficulty is above the sample median (harder exams). All regressions control for the standardized aggregate difficulty of the math exam, individual characteristics, and province and year fixed effects. Standard errors clustered at the province level are reported in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

TABLE 14. Nonlinear Effect of the First Question Difficulty

| | **Math Score** |
|---|---|
| **Reference group: Easy difficulty (below 50th percentile)** | |
| Normal difficulty (50th–90th percentile) | −0.171 |
| | (0.116) |
| Extreme difficulty (>90th percentile) | −0.513*** |
| | (0.121) |
| Difference: Extreme − Normal | −0.343*** |
| | (0.093) |
| Mean of DV | 0.000 |
| Observations | 13,648,306 |
| R-squared | 0.242 |
| Aggregate Difficulty Control | ✓ |
| Year FE | ✓ |
| Province FE | ✓ |
| Individual Control | ✓ |

*Notes*: This table examines nonlinearities in the effect of the first question difficulty in the math exam. All variables are standardized. The first question difficulty is divided into three categories based on the percentiles of its distribution: easy (< 50th percentile), normal ($\geq 50th, < 90th$ percentile), and extreme ($\geq$ 90th percentile). The first two rows report the coefficients on the normal- and extreme-difficulty indicators from a single regression, with the easy category as the reference group; each coefficient thus measures the effect relative to an easy first question. The row "Difference: Extreme − Normal" reports the difference between the two coefficients and its standard error. The regression controls for the standardized aggregate difficulty of the math exam, individual characteristics, and province and year fixed effects. Standard errors, clustered at the province level, are reported in parentheses.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

TABLE 15. Transmission Channels: Time Consuming and Cognitive Load

| | Math Score | | | |
|---|---|---|---|---|
| | (1) | (2) | (3) | (4) |
| *Panel A: Time Consuming V.S. Cognitive Load* | | | | |
| First Question Time Consuming | -0.072 | | -0.076 | |
| | (0.061) | | (0.056) | |
| First Question Cognitive Load | | -0.120*** | -0.124*** | |
| | | (0.037) | (0.038) | |
| R-squared | 0.236 | 0.237 | 0.239 | |
| *Panel B: Within Cognitive Load* | | | | |
| The Number of Knowledge Points | -0.103 | | | |
| | (0.067) | | | |
| The Abstract Level of Knowledge Points | | -0.137** | | |
| | | (0.048) | | |
| The Difficulty of Comprehend Question | | | -0.085** | |
| | | | (0.036) | |
| Innovative Level of Question | | | | -0.111** |
| | | | | (0.027) |
| R-squared | 0.236 | 0.238 | 0.236 | 0.237 |
| Mean of DV | 0.000 | 0.000 | 0.000 | 0.000 |
| Observations | 13,648,306 | 13,648,306 | 13,648,306 | 13,648,306 |
| Aggregate Difficulty Control | ✓ | ✓ | ✓ | ✓ |
| Year FE | ✓ | ✓ | ✓ | ✓ |
| Province FE | ✓ | ✓ | ✓ | ✓ |
| Individual Control | ✓ | ✓ | ✓ | ✓ |

Notes: This table examines two potential channels through which the salience of an unexpectedly difficult opening may translate into lower mathematics performance: time diverted from the remainder of the exam and cognitive interference with subsequent performance. Panel A compares calculation complexity, used as a proxy for the time required to solve the first question, with a composite measure of its cognitive demands. Panel B decomposes the composite measure into the number of knowledge points tested, the abstractness of those knowledge points, comprehension difficulty, and innovativeness. All variables in this table are standardized. All columns control for the aggregate difficulty of the math exam set, year and province fixed effects, and individual controls. Standard errors clustered at the province level are reported in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

# Online Appendix

# Blindsided at the Start: Evidence from LLM-Measured Exam Difficulty and Expanding Educational Inequality

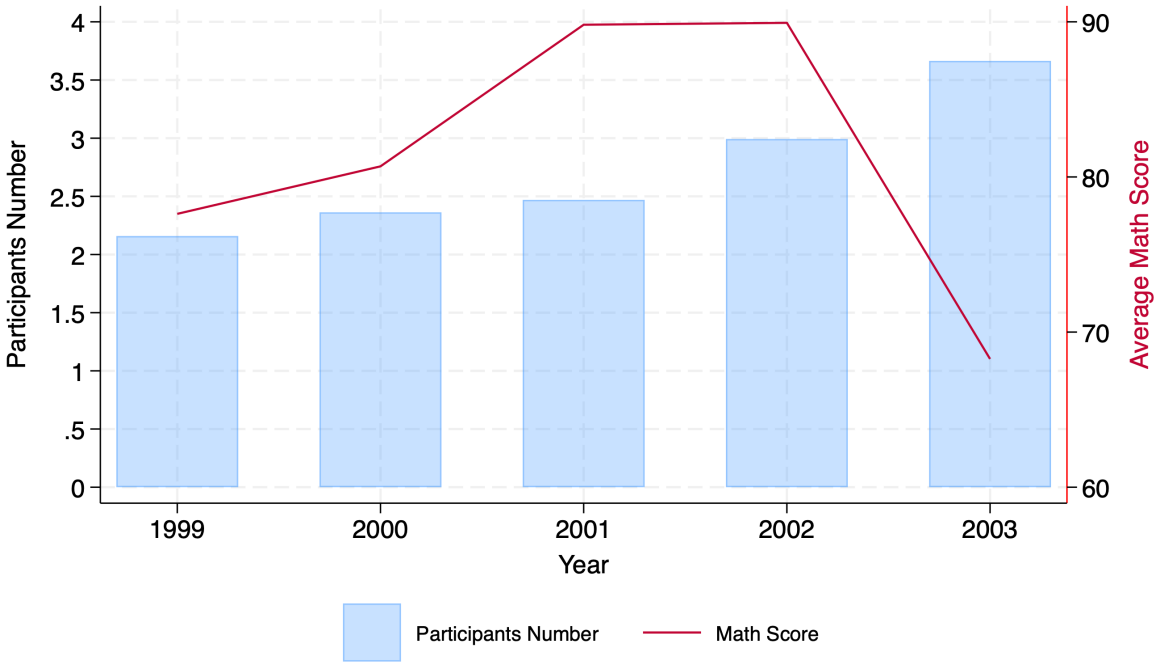
Yi Chen, Sikun Dou, Hongbin Li, Roujing Wu

August 2026

Yi Chen: ShanghaiTech University. Sikun Dou: Singapore Management University. Hongbin Li: Stanford University. Roujing Wu: University of Bologna. We would like to thank Benjamin Arold, Davide Cantoni, Hanming Fang, Henning Hermes, Vincenzo Scrutinio, and Weilong Zhang for suggestions that significantly improved the paper. This paper was presented at the University of Bologna, the University of Cambridge, and Ludwig Maximilian University of Munich. Chen acknowledges financial support from the National Science Foundation of China (72422019). All remaining errors are our own.

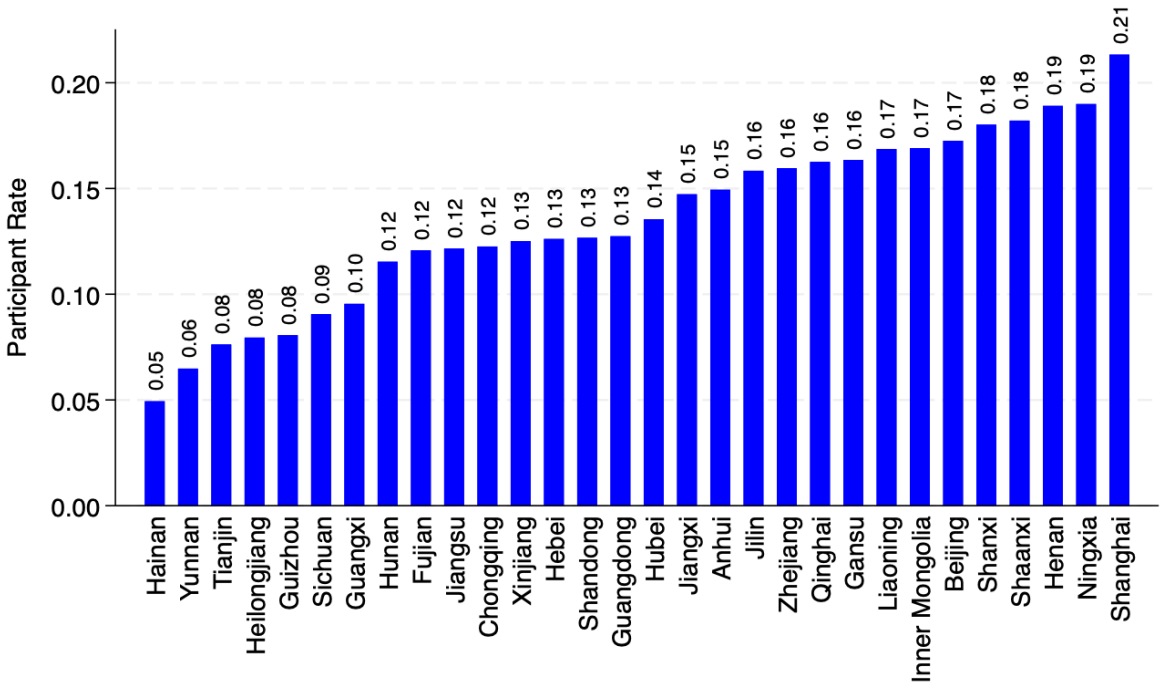
# Appendix A: Figures and Tables

Appendix Figure A1. Number of National College Entrance Exam Participants and Average Math Score



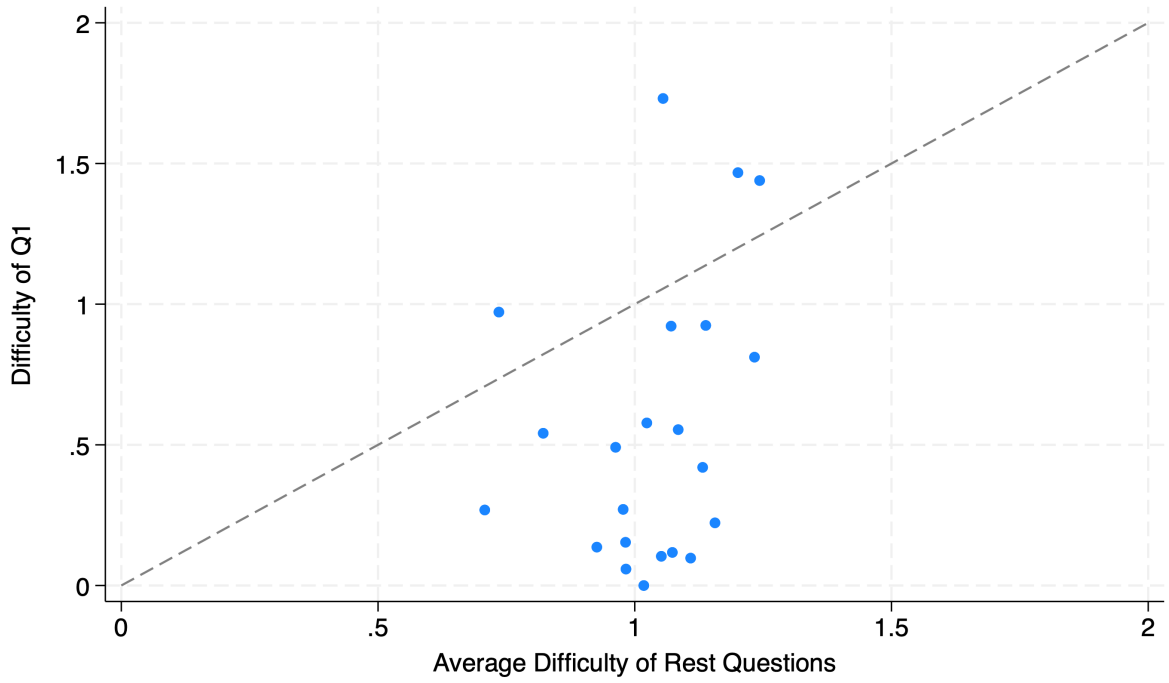

*Notes:* This figure exhibits the number of the NCEE participants and the average math score by year for the 1999–2003 exam cohorts in our sample (science and comprehensive tracks). Bars (left axis) show the number of exam participants, in millions. The solid line (right axis) shows the average raw math score, out of a total of 150 points.

Appendix Figure A2. National College Entrance Exam Participant Rates Across Provinces



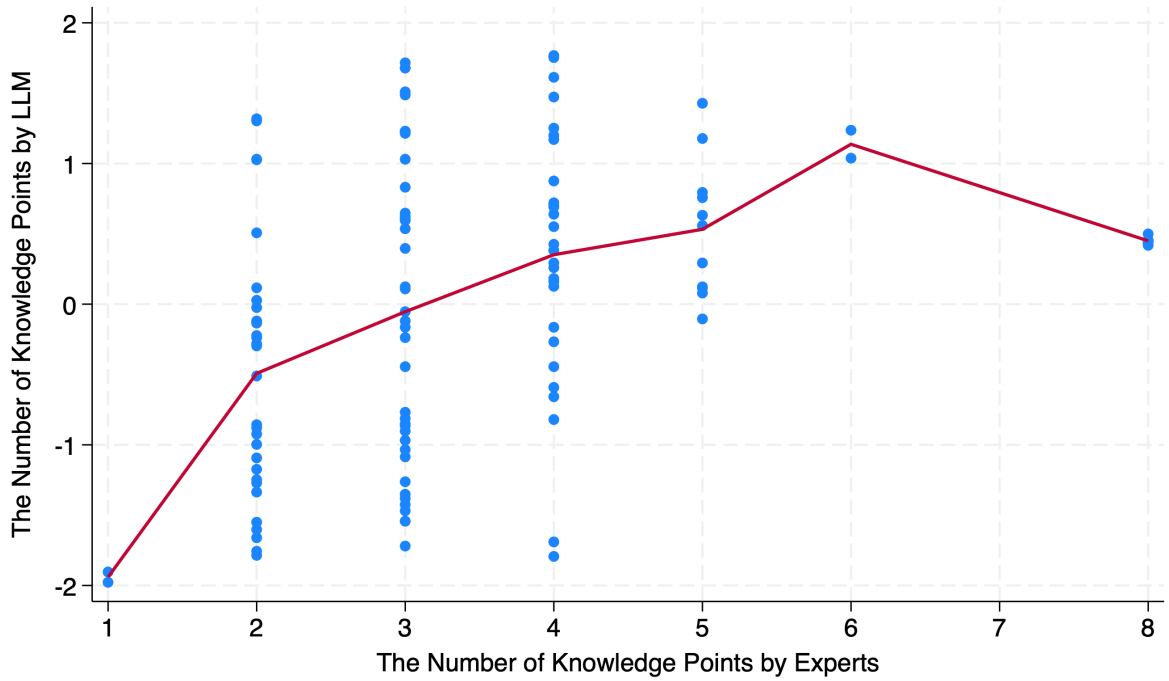

*Notes:* This figure reports the NCEE participation rate by province, averaged over the 1999–2003 exam cohorts in our sample (science and comprehensive track). The participation rate is defined as the number of NCEE participants divided by the population of the corresponding age cohort in each province (obtained from the 2000 National Population Census). Provinces are sorted in ascending order of the participation rate.

Appendix Figure A3. The First Question Difficulty versus Average Difficulty of Rest Questions



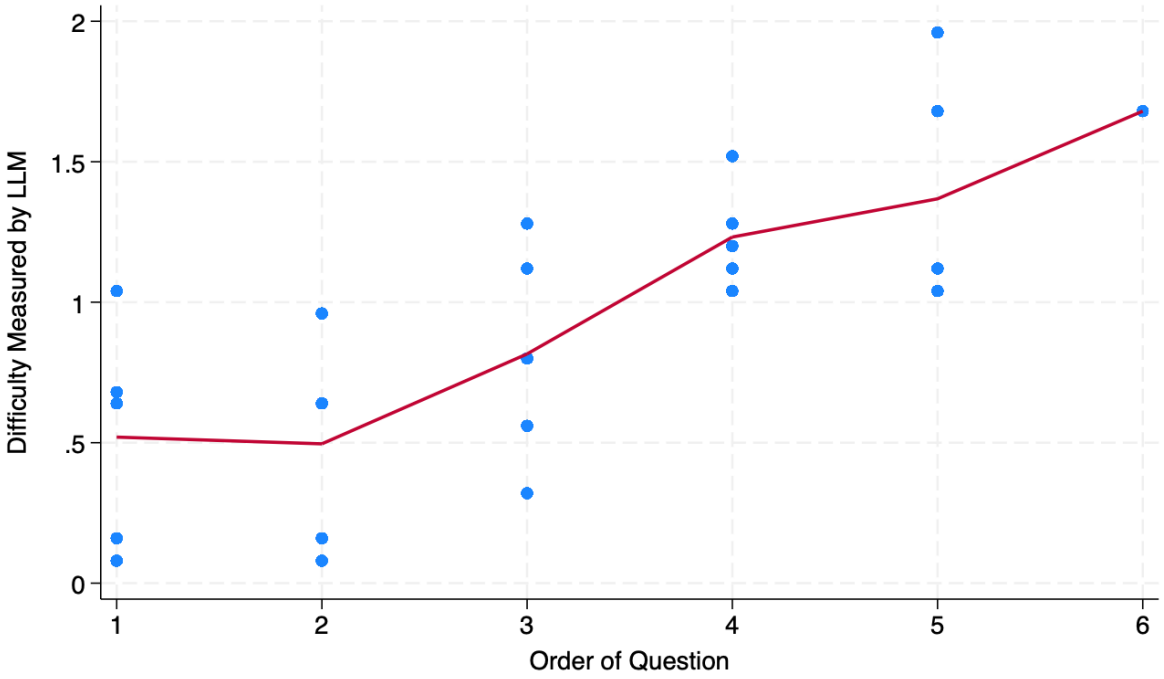

*Notes*: This figure plots the difficulty of the first question of the math exam (vertical axis) against the average difficulty of the remaining questions in the same exam set (horizontal axis). Each dot represents an exam set. The dashed line is the 45-degree line: dots below the line indicate exam sets in which the first question is easier than the average of the remaining questions. The correlation between the first question difficulty and the rest questions difficulty is 0.297 and statistically insignificant ($p$ = 0.180).

Appendix Figure A4. The Number of Knowledge Points: Expert versus LLM



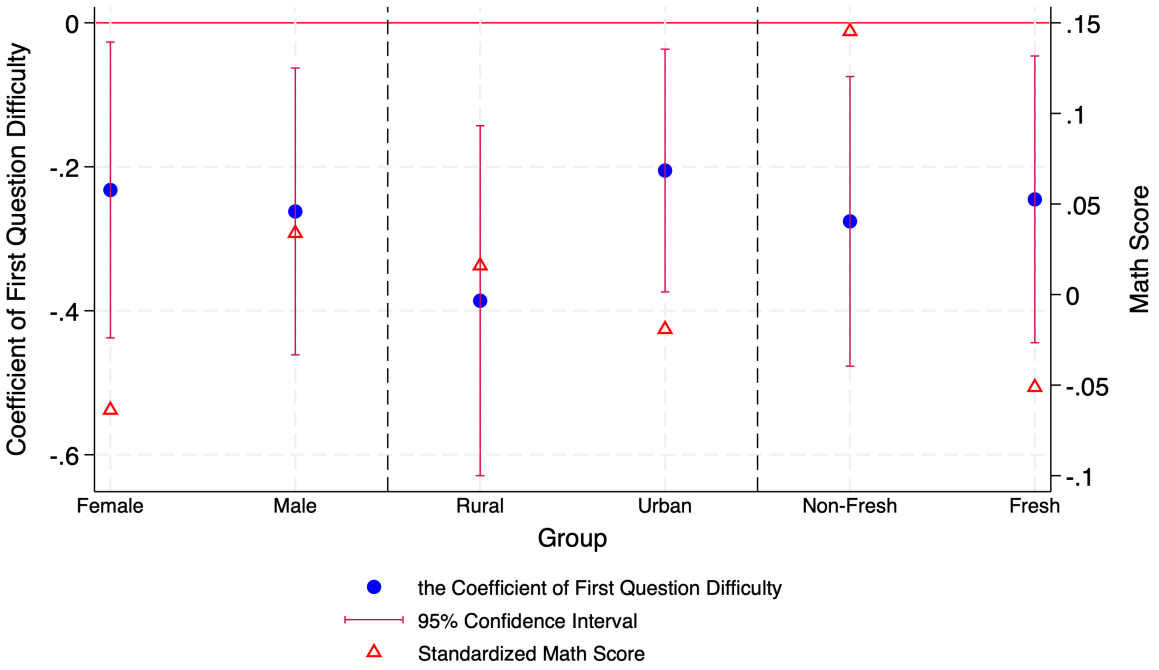

*Notes*: This figure compares the number of knowledge points tested by each question as counted by human experts (horizontal axis) with that measured by the ChatGPT-4o model (vertical axis). Each dot represents a problem-solving question in the math exam sets in our sample. The LLM-measured outcome is standardized. The solid line connects the average value of the LLM-based measure at each level of the expert count. The question-level correlation between the two measures is 0.368. The average LLM-based measure increases with the expert count (correlation is 0.802), indicating that the LLM-based measure captures the systematic variation in the number of knowledge points despite question-level noise.

Appendix Figure A5. The Order of Problem-Solving Question versus LLM-Measured Difficulty in 2026




*Notes*: This figure validates the LLM-measured raw difficulty of each problem-solving question in the 2026 NCEE math exam sets by its order of appearance in the exam. Each dot represents a problem-solving question; each exam set contains five or six problem-solving questions. The solid line connects the average difficulty at each position. NCEE questions are by design arranged in ascending order of difficulty. Consistent with this design, the LLM-measured difficulty increases with the question order: the question-level correlation between difficulty and position is 0.741, supporting the validity of the LLM-based difficulty measure. Since the 2026 exams were administered after the training data cutoff of the GPT-4o model, this validation is free from potential data contamination.

Appendix Figure A6. Heterogeneity Among Different Individual Characteristics




*Notes:* This figure presents heterogeneity in the association between exam difficulty and performance across individual characteristics. The horizontal axis shows six mutually exclusive groups: female, male, rural, urban, non-fresh graduates, and fresh graduates. For each group, the blue circle plots the estimated the coefficient on the difficulty of the first exam question from separate regressions of individual math scores. All regressions control for exam-level aggregate difficulty and include year and province fixed effects, but do not condition on individual covariates. The left vertical axis reports the magnitude of this coefficient; lower values indicate that first-question difficulty have more negative effect on math score for that group. Vertical red lines denote 95% confidence intervals based on standard errors clustered at the province level. Red triangles report the group-specific mean standardized math score and are read on the right vertical axis.

## Appendix Table A1. Summary Statistics

| Variable | Pre-Exam Application | | | Post-Exam Application | | | |
|---|---|---|---|---|---|---|---|
| | Observations | Mean | Std. Dev. | Observations | Mean | Std. Dev. | Difference |
| | **A. Demographic Information** | | | | | | |
| Female | 2,208,492 | .368 | .482 | 11,055,089 | .343 | .475 | .025 |
| Urban | 2,208,492 | .592 | .491 | 11,054,943 | .426 | .495 | .165*** |
| Fresh Graduate | 2,208,492 | .828 | .377 | 11,055,089 | .727 | .445 | .101*** |
| Science Track | 2,208,492 | .563 | .496 | 11,055,089 | .848 | .359 | -.286 |
| Comprehensive Track | 2,208,492 | .437 | .496 | 11,055,089 | .152 | .359 | .286 |
| Good High School | 2,015,767 | .271 | .445 | 10,142,708 | .250 | .433 | .021 |
| | **B. NCEE Performance** | | | | | | |
| NCEE Score | 2,208,474 | 417.124 | 88.120 | 11,054,865 | 423.252 | 96.221 | -6.128 |
| NCEE Rank | 2,208,492 | .501 | .289 | 11,055,089 | .500 | .289 | .001 |
| Chinese Score | 2,208,482 | 91.139 | 16.236 | 11,054,869 | 92.806 | 15.423 | -1.667 |
| Math Score | 2,208,492 | 83.646 | 22.786 | 11,054,808 | 80.422 | 25.817 | 3.224 |
| English Score | 2,208,492 | 81.890 | 21.560 | 11,053,771 | 80.359 | 22.822 | 1.531 |
| Comprehensive Subjects Score | 2,204,014 | 160.153 | 49.819 | 11,045,464 | 169.607 | 52.657 | -9.454 |
| | **C. College Admission Dummies** | | | | | | |
| Admitted | 2,208,492 | .712 | .453 | 11,055,089 | .597 | .491 | .115 |
| College Quality | 1,573,318 | .500 | .289 | 6,599,345 | .500 | .289 | .000 |
| STEM College | 2,208,492 | .156 | .363 | 11,055,089 | .153 | .360 | .003 |
| First Tier College | 2,208,492 | .141 | .348 | 11,055,089 | .123 | .328 | .018 |

*Notes*: This table reports summary statistics for the analysis sample, separately for pre-exam and post-exam application groups. All scores are reported in raw points: the *NCEE Score* is out of 750, the *Chinese Score*, *Math Score*, and *English Score* are each out of 150, and the *Comprehensive Subjects Score* is out of 300. *NCEE Rank* is the percentile rank of the NCEE score within the province–year, ranging from 0 to 1. The last column reports the difference in means between pre-exam and post-exam application groups, with standard errors clustered at the province level. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Appendix Table A2. Balance Check

| | **Demographic Characteristics** | | **Regional characteristics** | |
|---|---|---|---|---|
| | (1) | (2) | (3) | (4) |
| | Female | Urban | Per Capita GDP (log) | Admission Rate |
| **Panel A. Aggregate Difficulty** | | | | |
| Aggregate Difficulty | -0.005 | -0.005 | 0.073 | -0.013 |
| | (0.003) | (0.004) | (0.088) | (0.010) |
| R-squared | 0.013 | 0.013 | 0.993 | 0.819 |
| **Panel B. First Question Difficulty** | | | | |
| First Question Difficulty | -0.005 | 0.004 | -0.034 | -0.018 |
| | (0.006) | (0.006) | (0.101) | (0.011) |
| R-squared | 0.013 | 0.069 | 0.993 | 0.819 |
| Mean of DV | 0.347 | 0.453 | 9.591 | 0.617 |
| Observations | 13,648,730 | 13,648,584 | 141 | 142 |
| Year FE | ✓ | ✓ | ✓ | ✓ |
| Province FE | ✓ | ✓ | ✓ | ✓ |

*Notes*: This table shows balance checks on the LLM-measured exam difficulty. Each cell reports the coefficient from a separate regression of the characteristic listed in the column header on the standardized aggregate difficulty of the math exam (Panel A) or the standardized difficulty of its first question (Panel B), controlling for province and year fixed effects. The demographic characteristics (Female and Urban) are measured at the individual level; the regional characteristics (log per capita GDP in each county and college admission rate) are measured at the province–year level. The college admission rate is the ratio of admitted students to NCEE takers in the province–year. Standard errors clustered at the province level are reported in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Appendix Table A3. Spillover Effect of the First Question Difficulty

| | **Standardized Score** | | | | |
|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) |
| | Math | Chinese | English | Comprehensive | NCEE Score |
| First Question Difficulty | -0.181*** | -0.104 | -0.026 | -0.101 | -0.067** |
| | (0.055) | (0.085) | (0.030) | (0.082) | (0.030) |
| Mean of DV | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| Observations | 142 | 142 | 142 | 142 | 142 |
| R-squared | 0.870 | 0.403 | 0.884 | 0.724 | 0.935 |
| Aggregate Difficulty Control | ✓ | ✓ | ✓ | ✓ | ✓ |
| Year FE | ✓ | ✓ | ✓ | ✓ | ✓ |
| Province FE | ✓ | ✓ | ✓ | ✓ | ✓ |
| Demographic Characteristics Control | ✓ | ✓ | ✓ | ✓ | ✓ |

*Notes*: This table explores the spillover effects of the first question difficulty in the math exam on the performance in other subjects. Each column reports the coefficient on the standardized difficulty of the first math question from a separate regression at the province–year level, controlling for the standardized aggregate difficulty of the math exam, demographic characteristics, and province and year fixed effects. Standard errors clustered at the province level are reported in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Appendix Table A4. Effects of Exam Difficulty across the Math Score Distribution

| | Math Score | | | | | |
|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) |
| | 10% | 25% | 50% | 75% | 90% | Average |
| **Panel A. Aggregate Difficulty** | | | | | | |
| Aggregate Difficulty | -0.120** | -0.164*** | -0.212*** | -0.234*** | -0.228*** | -0.189** |
| | (0.048) | (0.051) | (0.045) | (0.035) | (0.025) | (0.040) |
| R-squared | 0.775 | 0.830 | 0.868 | 0.881 | 0.878 | 0.859 |
| **Panel B. First Question Difficulty** | | | | | | |
| First Question Difficulty | -0.165** | -0.188*** | -0.195*** | -0.194*** | -0.178*** | -0.181*** |
| | (0.063) | (0.063) | (0.057) | (0.050) | (0.054) | (0.055) |
| R-squared | 0.784 | 0.840 | 0.878 | 0.892 | 0.889 | 0.870 |
| Aggregate Difficulty Control | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Mean of DV | -1.361 | -0.652 | 0.057 | 0.714 | 1.278 | 0.000 |
| Observations | 142 | 142 | 142 | 142 | 142 | 142 |
| Year FE | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Province FE | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Demographic Characteristics Control | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |

*Notes*: This table presents the effects of exam difficulty across the distribution of the math score. The dependent variables are the 10th, 25th, 50th, 75th, and 90th percentiles and the average of the standardized math score, computed within each province–year. Panel A reports the coefficient on the standardized aggregate difficulty of the math exam; Panel B reports the coefficient on the standardized difficulty of its first question, controlling for the aggregate difficulty. All regressions are at the province–year level and control for demographic characteristics and province and year fixed effects. Standard errors, clustered at the province level are reported in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Appendix Table A5. Heterogeneous Effect of Aggregate Difficulty on Math Score

| | **Math Score** | | |
|---|---|---|---|
| | (1) | (2) | (3) |
| | Overall Sample | Pre-Exam Application | Post-Exam Application |
| Aggregate Difficulty | -0.183** | -0.155 | -0.124* |
| | (0.070) | (0.092) | (0.068) |
| Good High School | 0.615*** | 0.721*** | 0.599*** |
| | (0.031) | (0.052) | (0.033) |
| Aggregate Difficulty × Good High School | 0.003 | 0.023 | 0.003 |
| | (0.013) | (0.032) | (0.011) |
| Mean of DV | -0.006 | 0.122 | -0.005 |
| Observations | 12,523,569 | 2,015,767 | 10,142,335 |
| R-squared | 0.296 | 0.238 | 0.310 |
| Year FE | ✓ | ✓ | ✓ |
| Province FE | ✓ | ✓ | ✓ |
| Individual Control | ✓ | ✓ | ✓ |

*Notes*: This table examines whether the effect of the aggregate difficulty of the math exam on math performance varies by high school quality. All variable are standardized. Column (1) uses the full sample; Columns (2) and (3) split the sample into pre-exam and post-exam application groups, respectively. All regressions control for individual characteristics and province and year fixed effects. Standard errors clustered at the province level are reported in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Appendix Table A6. Whether the Ability Matters for the Negative Shock

| | **Math Score** | | |
|---|---|---|---|
| | (1) | (2) | (3) |
| | Overall Sample | Pre-Exam Application | Post-Exam Application |
| First Question Difficulty | -0.250*** | -0.219*** | -0.239*** |
| | (0.127) | (0.043) | (0.079) |
| Top 50% Students | 0.933*** | 0.857*** | 0.946*** |
| | (0.127) | (0.084) | (0.021) |
| First Question Difficulty × Top 50% Students | -0.021 | 0.021 | -0.041 |
| | (0.035) | (0.046) | (0.032) |
| Mean of DV | -0.006 | 0.122 | -0.005 |
| Observations | 13,458,309 | 2,144,216 | 10,943,658 |
| R-squared | 0.627 | 0.605 | 0.638 |
| Year FE | ✓ | ✓ | ✓ |
| Province FE | ✓ | ✓ | ✓ |
| High School FE | ✓ | ✓ | ✓ |
| Individual Control | ✓ | ✓ | ✓ |

*Notes*: This table examines whether the effect of first-question difficulty on math performance varies with student ability. Top 50% Students is an indicator equal to one if the student ranked in the top half within his/her own high school-province-year-track cell based on aggregate NCEE score. First Question Difficulty is standardized to have zero mean and unit variance; the dependent variable is the standardized math score. The coefficient on First-Question Difficulty therefore captures the effect for students in the bottom half of the achievement distribution, while the interaction coefficient measures the differential effect for students in the top half. Column (1) reports estimates for the full sample. Columns (2) and (3) restrict the sample to province-years in which university preferences were submitted before and after the examination, respectively. All regressions control for individual characteristics and school, province and year fixed effects. Standard errors clustered at the province level are reported in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

## Appendix Table A7. Heterogeneous Effect in Each Decile of Aggregate NCEE Score

| | Math Score | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) | |
| | 10% | 20% | 30% | 40% | 50% | 60% | 70% | 80% | 90% | 100% |
| First Question Difficulty | -0.183 | -0.248* | -0.288** | -0.301** | -0.310*** | -0.310*** | -0.314*** | -0.305*** | -0.302*** | -0.289*** |
| | (0.120) | (0.126) | (0.118) | (0.111) | (0.106) | (0.099) | (0.093) | (0.090) | (0.079) | (0.072) |
| Good High School | 0.014 | 0.006 | 0.008 | 0.006 | 0.010 | 0.013 | 0.014* | 0.015 | 0.022* | 0.085*** |
| | (0.025) | (0.012) | (0.012) | (0.011) | (0.009) | (0.008) | (0.008) | (0.010) | (0.012) | (0.017) |
| First Question Difficulty × Good High School | 0.026 | 0.051** | 0.062*** | 0.062*** | 0.062*** | 0.059*** | 0.057*** | 0.058*** | 0.056** | 0.060** |
| | (0.022) | (0.020) | (0.021) | (0.022) | (0.022) | (0.022) | (0.020) | (0.021) | (0.022) | (0.026) |
| Mean of DV | -1.402 | -0.792 | -0.467 | -0.225 | -0.021 | 0.162 | 0.340 | 0.521 | 0.745 | 0.141 |
| Observations | 1,253,138 | 1,256,698 | 1,252,776 | 1,245,622 | 1,241,121 | 1,244,123 | 1,244,036 | 1,249,595 | 1,258,742 | 1,277,718 |
| R-squared | 0.274 | 0.435 | 0.497 | 0.532 | 0.553 | 0.570 | 0.582 | 0.582 | 0.582 | 0.479 |
| Aggregate Difficulty | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Year FE | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Province FE | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Demographic Characteristics Control | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |

*Notes*: This table reports heterogeneous effects across deciles of the aggregate NCEE score distribution — as a proxy of student's ability . The sample is divided into ten mutually exclusive groups based on students' total NCEE scores, with columns (1)–(10) corresponding to the first (bottom) through tenth (top) deciles, respectively. The dependent variable is the standardized math score. First Question Difficulty denotes the standardized difficulty of the first question on the math exam, and Good High School is an indicator for attending a good high school. The interaction term captures whether the effect of first-question difficulty differs for students attending good high schools. All regressions are at control for aggregate difficulty, demographic characteristics and province and year fixed effects. Standard errors, clustered at the province level are reported in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Appendix Table A8. Robustness Test on Restricted Sammple

| | (1) | (2) | (3) | (4) | (5) | (6) |
|---|---|---|---|---|---|---|
| | Math Score | NCEE Score | NCEE Rank | Admitted | Uni Quality | STEM Uni |
| **Panel A. Sample Excludes First-Time Adopters of the New Exam** | | | | | | |
| First Question Difficulty | -0.297** | -0.216*** | -0.000 | -0.019 | -0.001*** | -0.001 |
| | (0.117) | (0.038) | (0.001) | (0.021) | (0.000) | (0.043) |
| Mean of DV | 0.019 | 0.003 | 0.500 | 0.605 | 0.500 | 0.251 |
| Observations | 12,118,498 | 12,118,473 | 12,118,675 | 12,113,602 | 7,332,420 | 7,332,420 |
| R-squared | 0.239 | 0.138 | 0.015 | 0.083 | 0.011 | 0.043 |
| **Panel B. Sample Excludes Provinces with Province-Specific Exam Set** | | | | | | |
| First Question Difficulty | -0.270** | -0.166*** | -0.001* | -0.025 | -0.000 | -0.006 |
| | (0.113) | (0.060) | (0.000) | (0.021) | (0.000) | (0.009) |
| Mean of DV | 0.018 | 0.013 | 0.500 | 0.633 | 0.500 | 0.156 |
| Observations | 11,432,761 | 11,432,801 | 11,433,038 | 11,427,965 | 7,229,973 | 7,229,973 |
| R-squared | 0.252 | 0.132 | 0.015 | 0.077 | 0.009 | 0.043 |
| Aggregate Difficulty Control | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Year FE | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Province FE | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Individual Control | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |

*Notes*: This table reports robustness checks based on restricted samples and updated prompt for measurement. Panel A excludes province–years in which the new exam set is administered for the first time, to address the concern that unfamiliarity with a new exam format, rather than first-question difficulty per se, drives the results. Panel B excludes province–years in which the NCEE math exam is province-specific; the mapping between province–years and exam sets is reported in Table 1. Since students in the comprehensive track are exactly those taking the Guangdong exam sets (Exam Sets 4, 8, 13, and 18), the exclusion in Panel B also removes all comprehensive-track students, leaving a sample of science-track students only. All regressions control for the standardized aggregate difficulty of the math exam, individual characteristics, and province and year fixed effects. Standard errors clustered at the province level are reported in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

## Appendix Table A9. Robustness Test for Alternative Measures of Difficulty

| | **Baseline** | | | | | | **Heterogeneity** | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) | (10) | (11) | (12) |
| | Math Score | NCEE Score | NCEE Rank | Admitted | Uni Quality | STEM Uni | Math Score | NCEE Score | NCEE Rank | Admitted | Uni Quality | STEM Uni |
| **Panel A. Alternative Prompt for Measuring Difficulty** | | | | | | | | | | | | |
| First Question Difficulty | -0.224*** | -0.199*** | -0.001 | -0.022 | -0.001** | -0.009 | -0.264*** | -0.210*** | -0.005 | -0.028 | -0.007 | -0.007 |
| | (0.080) | (0.047) | (0.001) | (0.018) | (0.000) | (0.008) | (0.084) | (0.048) | (0.004) | (0.020) | (0.004) | (0.009) |
| Good High School | | | | | | | 0.682*** | 0.777*** | 0.254 | 0.173*** | 0.209*** | 0.092*** |
| | | | | | | | (0.041) | (0.051) | (0.016) | (0.016) | (0.012) | (0.006) |
| First Question Difficulty × Good High School | | | | | | | 0.062* | 0.001 | 0.007 | -0.009 | 0.011 | 0.004 |
| | | | | | | | (0.031) | (0.033) | (0.011) | (0.012) | (0.010) | (0.004) |
| R-squared | 0.241 | 0.130 | 0.015 | 0.084 | 0.010 | 0.043 | 0.303 | 0.230 | 0.136 | 0.109 | 0.097 | 0.051 |
| **Panel B. Alternative LLM Model for Measuring Difficulty: Claude Sonnet 4.6** | | | | | | | | | | | | |
| First Question Difficulty | -0.306** | -0.160 | -0.001 | -0.044*** | -0.000 | -0.030 | -0.354** | -0.188 | -0.010* | -0.040** | -0.014*** | -0.030* |
| | (0.146) | (0.118) | (0.001) | (0.016) | (0.001) | (0.015) | (0.154) | (0.119) | (0.005) | (0.018) | (0.003) | (0.016) |
| Good High School | | | | | | | 0.709*** | 0.814*** | 0.270*** | 0.142*** | 0.232*** | 0.092*** |
| | | | | | | | (0.040) | (0.055) | (0.019) | (0.027) | (0.010) | (0.007) |
| First Question Difficulty × Good High School | | | | | | | 0.070** | 0.029 | 0.018 | -0.030* | 0.026*** | 0.003 |
| | | | | | | | (0.029) | (0.035) | (0.012) | (0.017) | (0.009) | (0.005) |
| R-squared | 0.238 | 0.129 | 0.015 | 0.084 | 0.010 | 0.043 | 0.300 | 0.228 | 0.136 | 0.109 | 0.098 | |
| Mean of DV | 0.000 | 0.000 | 0.500 | 0.612 | 0.500 | 0.248 | 0.000 | 0.000 | 0.500 | 0.612 | 0.500 | 0.248 |
| Observations | 13,648,306 | 13,648,342 | 13,648,584 | 13,643,511 | 8,347,703 | 8,347,703 | 12,523,569 | 12,523,600 | 12,523,818 | 12,523,818 | 7,634,228 | 7,634,228 |
| Aggregate Difficulty Control | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Year FE | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Province FE | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Individual Control | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |

*Notes*: This table reports robustness tests using alternative measures of question difficulty. In Panel A, question difficulty is assessed by GPT-4o, the same LLM used in the baseline analysis, using an alternative prompt that instructs the model to evaluate each question from the perspective of a typical Chinese high-school student during 1999–2003 and to rely only on knowledge covered by the Chinese high-school curriculum. In Panel B, question difficulty is assessed by Claude Sonnet 4.6 using the same prompt as in the baseline analysis. All regressions control for the standardized aggregate difficulty of the mathematics examination, individual characteristics, and province and year fixed effects. All regressions control for the standardized aggregate difficulty of the math exam, individual characteristics, and province and year fixed effects. Standard errors clustered at the province level are reported in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Appendix Table A10. Robustness Test For Different Thresholds on Good High School

| | (1) | (2) | (3) | (4) | (5) | (6) |
|---|---|---|---|---|---|---|
| | Math Score | NCEE Score | NCEE Rank | Admitted | Uni Quality | STEM Uni |
| **Panel A. Good High School— Above Average Level** | | | | | | |
| First Question Difficulty | -0.3111*** | -0.238*** | -0.012** | -0.035 | -0.013*** | -0.012 |
| | (0.100) | (0.068) | (0.004) | (0.022) | (0.004) | (0.011) |
| Good High School | 0.600*** | 0.745*** | 0.237*** | 0.172*** | 0.192*** | 0.084*** |
| | (0.028) | (0.031) | (0.010) | (0.009) | (0.009) | (0.005) |
| First Question Difficulty × Good High School | 0.076*** | 0.038 | 0.014 | -0.006 | 0.015 | 0.004 |
| | (0.027) | (0.027) | (0.009) | (0.011) | (0.009) | (0.004) |
| R-squared | 0.311 | 0.242 | 0.152 | 0.111 | 0.104 | 0.051 |
| **Panel B. Good High School— Above** 90% | | | | | | |
| First Question Difficulty | -0.291*** | -0.225*** | -0.005* | -0.034 | -0.005** | -0.009 |
| | (0.099) | (0.069) | (0.003) | (0.021) | (0.002) | (0.010) |
| Good High School | 0.733*** | 0.906*** | 0.285*** | 0.212*** | 0.225*** | 0.103*** |
| | (0.028) | (0.030) | (0.011) | (0.011) | (0.008) | (0.007) |
| First Question Difficulty × Good High School | 0.112*** | 0.082*** | 0.015* | -0.003 | 0.016** | 0.004 |
| | (0.032) | (0.025) | (0.008) | (0.012) | (0.008) | (0.004) |
| R-squared | 0.282 | 0.195 | 0.094 | 0.101 | 0.072 | 0.049 |
| Mean of DV | -0.006 | -0.001 | 0.501 | 0.610 | 0.503 | 0.249 |
| Observations | 12,523,569 | 12,523,600 | 12,523,818 | 12,523,818 | 7,634,228 | 7,634,228 |
| Aggregate Difficulty Control | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Year FE | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Province FE | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Individual Control | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |

*Notes:* This table examines the robustness of the heterogeneity results to alternative definitions of a good high school. In Panel A, a high school is classified as good if its first-tier rate is above the average level; in Panel B, students graduate from the good high school if its first-tier rate above the 90th percentile. Each column reports a separate regression of the dependent variable listed in the column header on first-question difficulty, the good high school indicator, and their interaction. Columns (5) and (6) are estimated on the sample of admitted students. All regressions control for the standardized aggregate difficulty of the math exam, individual characteristics, and province and year fixed effects. Standard errors clustered at the province level are reported in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Appendix Table A11. Post-Exam Application: The Effect of the First Question Difficulty on Performance and Admission Outcome

| | **Baseline** | | | | | | **Heterogeneity** | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) | (10) | (11) | (12) |
| | Math Score | NCEE Score | NCEE Rank | Admitted | Uni Quality | STEM Uni | Math Score | NCEE Score | NCEE Rank | Admitted | Uni Quality | STEM Uni |
| First Question Difficulty | -0.285*** | -0.280*** | -0.003 | -0.054** | -0.003 | -0.020 | -0.299*** | -0.285*** | -0.007 | -0.050* | -0.012** | -0.022 |
| | (0.088) | (0.100) | (0.004) | (0.026) | (0.004) | (0.020) | (0.095) | (0.098) | (0.006) | (0.027) | (0.005) | (0.020) |
| Good High School | | | | | | | 0.606*** | 0.758*** | 0.242*** | 0.179*** | 0.193*** | 0.088*** |
| | | | | | | | (0.031) | (0.033) | (0.011) | (0.010) | (0.009) | (0.006) |
| First Question Difficulty × Good High School | | | | | | | 0.053 | 0.017 | 0.014 | -0.016 | 0.018 | 0.001 |
| | | | | | | | (0.033) | (0.035) | (0.012) | (0.014) | (0.012) | (0.005) |
| Mean of DV | -0.005 | 0.025 | 0.500 | 0.597 | 0.500 | 0.256 | -0.013 | 0.021 | 0.501 | 0.594 | 0.502 | 0.257 |
| Observations | 10,142,335 | 10,142,372 | 10,142,584 | 10,142,584 | 6,026,784 | 6,026,784 | 10,142,335 | 10,142,372 | 10,142,584 | 10,142,584 | 6,026,784 | 6,026,784 |
| R-squared | 0.256 | 0.148 | 0.019 | 0.084 | 0.011 | 0.039 | 0.314 | 0.240 | 0.134 | 0.106 | 0.094 | 0.047 |
| Aggregate Questions Difficulty Control | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Year FE | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Province FE | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Individual Control | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |

*Notes*: This table shows the effects of the first question difficulty in the NCEE performance and admission outcomes, restricting the sample to post-exam application. Columns (1)–(6) report the baseline effects; Columns (7)–(12) examine heterogeneity by high school quality. All regressions control for the standardized aggregate difficulty of the math exam, individual characteristics, and province and year fixed effects. Standard errors clustered at the province level are reported in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

# Appendix B: Methodology — Measuring Question Difficulty

## B.1. The Measurement Challenge

Our research design requires a measure of question difficulty that is comparable across provinces and over time. Realized scores and score-based item statistics do not satisfy this requirement because of institutional features of the NCEE.

Realized scores depend on both test difficulty and the composition of the test-taking population. In our setting, variation along both dimensions makes them an unreliable measure of difficulty. Over time, the higher-education expansion launched in 1999 transformed the examinee pool. As a result, changes in scores mix changes in exam difficulty with changes in exam takers. Cross-sectional comparability is also limited. NCEE participation rates in our data range from 5 percent of the relevant cohort in Hainan to 21 percent in Shanghai.[1] In the language of psychometrics, scores from different test forms taken by non-equivalent groups are not comparable without formal equating, which requires either common anchor items across forms or randomly equivalent examinee groups (Holland and Dorans 2006; Kolen and Brennan 2014). The NCEE, whose papers are deliberately differentiated across regions, satisfies neither condition.

The solution to this confounding is item response theory (IRT), which models the probability of a correct response as a function of latent examinee ability and item-level parameters, thereby separating what a question demands from who happens to attempt it (Rasch 1960; Lord and Novick 1968; Lord 1980; Hambleton and Swaminathan 1985). IRT is infeasible here, for two reasons. First, identification of item parameters—in the Rasch model as in its two- and three-parameter generalizations—requires the full person-by-item response matrix, that is, each examinee's score on each individual question (van der Linden and Hambleton 1997; Embretson and Reise 2000). No such

[1] See Figure A2.

data exist for the NCEE: administrative records preserve only subject-level total scores, and item-level answer sheets are never released. This is not an idiosyncratic limitation of our data but a generic feature of large-scale, high-stakes examinations, in which security and confidentiality protocols preclude the release of item-level microdata; as Bruhn et al. (2025) note, most administrative settings observe only total scores, which compress the underlying response information. Second, even were such data available, placing item parameters from different papers on a common scale would again require common items or randomly equivalent groups across provinces (Kolen and Brennan 2014). Measuring difficulty for the NCEE therefore requires an approach that operates on the content of the questions themselves rather than on the responses of those who answered them.

We propose such an approach: we use a large language model (LLM) to construct a difficulty index directly from the digitized text of each question. The method builds on a growing literature that applies machine learning and LLMs to educational measurement, problem solving and analysis (OpenAI 2023; Benedetto et al. 2023; Yuan et al. 2023; Zhang et al. 2023), and on the rapidly expanding use of LLMs as measurement instruments in economics (Chen et al. 2024; Arold et al. 2025; Rodrik and Geyiktepe 2026). Because the index is constructed ex ante from item content, it is, by design, invariant to the size and composition of the examinee pool. This invariance is the key requirement that score-based measures fail to satisfy.

### B.2. GPT Measure of Math Difficulty

We construct the difficulty index from a pairwise tournament among questions. Each question is compared repeatedly with other questions of the same track and the same type of questions, both within its own exam set and across sets; the cross-set comparisons are essential, as they place all 484 questions on a single common scale. For each

pair, we instruct the model to judge whether the first question is more difficult than, easier than, or roughly comparable in difficulty to the second. We access ChatGPT-4o model through the API with the temperature parameter set to zero, so that the model returns its modal judgment and the procedure is deterministic. The prompt used for this evaluation is as follows:

> *You are expert on education and you are able to compare the difficulties of two questions. Please compare the difficulties of these two questions, directly tell me the conclusion: [question 1], [question 2]. Please give the conclusion: 1. question 1 is a bit more difficult; 2. question 2 is a bit more difficult; 3. the difficulties are roughly equal.*

We elicit each pairing twice, once in each order, so that any position bias cancels out in the aggregate (Zheng et al. 2023). Each comparison awards a question two points for a win, one point for a tie, and zero points for a loss. Question-level difficulty, $d_{ej}$, is the question's average score across all comparisons in which it appears and thus ranges from zero to two. Because LLMs are poorly calibrated on numerical rating tasks, we elicit only ordinal judgments and aggregate them, following the classical paired-comparison literature (Bradley and Terry 1952) and standard practice in LLM evaluation (Zheng et al. 2023).

Exam-set difficulty is the point-weighted average of question difficulties:

$$D_e = \sum_j \left( \frac{s_{ej}}{150} \right) d_{ej}, \tag{3}$$

where $s_{ej}$ denotes the point value of question $j$ on exam set $e$ and $\sum_j s_{ej} = 150$. This weighting makes the aggregate index reflect the contribution of each item to the formal scoring rule. The baseline regressions use the standardized values of $d_{e1}$ and $D_e$.

### B.3. Auxiliary Content Measures

For the mechanism analysis, we apply the same paired-comparison procedure to five additional dimensions. Calculation complexity measures the length and complexity of the operations needed to obtain a solution and serves as a proxy for time requirements. The cognitive-demand measure is based on four components: the number of knowledge points required, the abstractness of those knowledge points, the difficulty of comprehend the question, and the question's innovativeness relative to standard curricular material. The comparison prompt instructs the model to adopt the perspective of a typical Chinese high school student in 1999–2003 and to use only the curriculum and methods available during that period.

We standardize the four component measures and use their unweighted mean as the composite cognitive-demand index. The exact prompt is reproduced below:

> *You are expert on Chinese high-school mathematics education and exam design, and you are able to compare the difficulties of two questions along SIX dimensions.*
>
> *You must stand at the perspective of a typical Chinese high-school student taking exams around 1999–2003, and you can ONLY use knowledge and techniques from the official "Chinese high-school curriculum and exam syllabi in that period (no college-level methods)." Compare the two questions on these SIX dimensions:*

1) *more_knowledge_points: which question involves more distinct mathematical concepts/knowledge points;*
2) *higher_abstract_level: which question relies on more abstract or conceptually demanding ideas;*
3) *harder_to: which question is harder to understand or interpret correctly;*
4) *more_innovative: which question is more non-routine/innovative relative to 1999–2003 exams;*

5) *more_complex_calculation: which question has more complex/lengthy calculations;*

6) *more_difficult: which question is overall harder for a typical 1999–2003 student.*

(a) *[1:] question 1 is a bit higher on this dimension;*

(b) *[2:] question 2 is a bit higher on this dimension;*

(c) *[0:] the two questions have rougly equal level on this dimension.*

*Your FINAL ANSWER must be EXACTLY SIX lines, each in the following form:*
*more_knowledge_points: [0 or 1 or 2];*
*higher_abstract_level: [0 or 1 or 2];*
*harder_to_understand:[0 or 1 or 2]; more_innovative: [0 or 1 or 2];*
*more_complex_calculation: [0 or 1 or 2];*
*more_difficult: [0 or 1 or 2].*
*Do NOT output any other text.*

We also use this prompt for remeasure the difficulty of questions and do the robustness check in section 6.

Two practical properties of the procedure are worth noting. It is inexpensive, with a marginal cost that is trivial relative to expert grading of nearly five hundred questions; [2] and it is fully replicable, in that any researcher supplying the same prompt, model, and temperature recovers the identical index.

### B.4. Measurement Results

Figure 1 plots the estimated difficulty of every question in each of the 22 exam sets by different question types in our sample, ordered by the question's position within the

[2] It costs around 123 dollars and 6 hours under GPT-4o model; For Claude Sonnet 4.6 model, it takes around 148.1 dollars and 7 hours.

exam; Table 3 reports aggregate and first-question difficulty and corresponding ranks for each set. Difficulty is reported in raw, unstandardized units on a 0–2 scale, with the solid line in each panel of Figure 1 marking the exam set's average difficulty and dash line shows the average difficulty within each question type. Item difficulty varies substantially within each exam set and type.

Difficulty in exam set level ranges from 0.693 (2002 Old standard) to 1.245 (2002 Beijing). Notably, this variation is as large within years as between them: students sitting the "same" NCEE in the same year, but in different provinces, faced markedly different tests—in 2003 alone, aggregate difficulty spans from 0.954 (Beijing) to 1.219 (Old standard). Nor does the variation reflect persistent traits of provinces or exam paper series: the old-standard national paper, for instance, ranks among the easiest sets in 2002 (0.693) but the hardest in 2003 (1.219). Difficulty is thus best understood as a exam set-specific shock rather than a fixed characteristic of the administering jurisdiction, which is precisely the variation our empirical strategy requires.

The difficulty of the first question is central to our research design. As Figure 1 shows, the first question in most exam sets falls visibly below the exam set's and type's average difficulty line, and the data confirm this pattern quantitatively: the mean difficulty of first questions is 0.447, less than half the mean of 0.979 across all questions, and in 19 of the 22 exam sets the opening question is easier than the average question on its own paper. This is consistent with the conventional practice, noted in subsection 2.3, of opening a math exam with an easy question, and the convention is sufficiently regular that a student trained on past exams would rationally expect an easy start—it is precisely this convention that gives an unexpectedly difficult first question its interpretation as a salient, reference-violating shock rather than merely one more hard question among many.

Despite this general pattern, first-question difficulty is only weakly related to overall exam difficulty. Across the 22 exam sets, a regression of first-question difficulty on ag-

gregate difficulty yields a coefficient of 0.394. The relationship is weaker when aggregate difficulty is measured using the remaining questions: the correlation is 0.297. Figure A3 illustrates this limited relationship, showing that exam sets with similar average difficulty across the remaining questions can begin with substantially different levels of difficulty. Thus, first-question difficulty captures variation that is distinct from overall exam difficulty. Our main specifications control for aggregate exam difficulty, so the effect of first-question difficulty is identified from differences in the opening question across exams with comparable overall difficulty.

In addition to the difficulty of the first question, we construct two auxiliary measures that capture distinct dimensions of the first question: time required to complete the question and cognitive load. Although these measures are positively correlated, they are not interchangeable. Figure 3 presents the first-question indexes across exam sets. The correlation between difficulty and time consumption is 0.587, while that between difficulty and cognitive load is 0.692. The correlation between time consumption and cognitive load is 0.007. We use the difficulty index as our baseline measure and draw on the two auxiliary measures to investigate the channels through which the difficulty of the opening question may affect performance in math.

### B.5. Validation

We validate the index in three ways: against the sequence of questions, against expert human judgment, and against the realized performance of actual exam takers. To address concerns about potential training data contamination, we further validate our results using contemporary exam questions.

*Question sequence.* NCEE exam sets are constructed, by explicit design convention, so that difficulty rises with question position within each type of question. This design feature provides a natural validation test: although our measure is constructed without

information on question position, it should recover the systematic relationship between question order and difficulty implied by the exam design. The results confirm this prediction. Scaling the position of the six problem-solving questions from 0 (first) to 100 (last), average measured difficulty increases monotonically in position shown in Figure 2 Panel A, the correlation is 0.952.

*Expert judgment.* We compare the difficulty index with published assessments by NCEE mathematics educators, coded on a six-point scale from "easy" (0) to "difficult" (100) in increments of 20. Figure 2 presents that LLM-measured difficulty is monotonically increasing in the expert rating across the entire scale questions rated "easy" by experts receive an average index near the bottom of the distribution, those rated "difficult" near the top. The correlation of is 0.963 across expert-rating averages. [3]

*Math Score.* If the index captures true difficulty, mechanically students should have lower math score. We estimate

$$\text{MathScore}_{i,p,t} = \beta_0 + \beta_1 D_e + X_i + \gamma_t + \gamma_p + \varepsilon_p, \tag{4}$$

where $\text{MathScore}_{i,p,t}$ is the standardized math score of student $i$ in province $p$ and year $t$, $D_e$ is the standardized difficulty of the exam set the student faced, $X_i$ contains individual controls (female, urban and fresh graduates), and $\gamma_t$ and $\gamma_p$ are year and province fixed effects; standard errors are clustered at the province level. Identification of $\beta_1$ captures the effect of exam set difficulty on math scores. As shown in Table 4, a one-standard-deviation increase in aggregate difficulty reduces math scores by 0.189 standard deviations in the most demanding specification with both fixed effects and individual controls.

[3] Validation of the number of knowledge points is shown in Figure A4. The average LLM-based measure increases with expert count, and corresponding correlation is 0.802.

*Independent Validation.* Finally, to address concerns about potential training-data contamination, we apply the same LLM (ChatGPT 4o) and identical prompt to the 2026 NCEE mathematics examinations. These examinations postdate the model's training data and therefore provide a setting in which the assessment is unlikely to be driven by prior exposure to the exam questions. The resulting index ranks Shanghai (1.156) above Beijing (1.039), National Paper I (0.949) well above National Paper II (0.789), and easiest one is Tianjin (0.781)—orderings that are consistent with the prevailing views of educators and contemporaneous public commentary following the examination. As shown in Figure A5, the within-paper position gradient in 2026 is likewise upward sloping. Taken together, these results show that the difficulty index continues to capture meaningful variation in exam difficulty in a setting that is largely insulated from concerns about training-data contamination.

Importantly, the ability of the same model to recover these patterns in post-training examinations suggests that the performance of the difficulty index is not driven solely by prior exposure to the historical questions. If the index were primarily capturing memorized information about the 1999–2003 examinations, we would not expect it to generalize to questions that were not available during model training. The 2026 results therefore provide evidence that the index captures generalizable features of mathematical difficulty rather than question-specific memorization.

Having established that our measure tracks question difficulty in a manner consistent with external benchmarks, we next describe how it varies across exam sets and question positions in our estimation sample, and how we merge it with administrative student records section 3.