UNIVERSIDADE DE SÃO PAULO
FACULDADE DE ECONOMIA, ADMINISTRAÇÃO E CONTABILIDADE
DEPARTAMENTO DE ECONOMIA
PROGRAMA DE PÓS-GRADUAÇÃO EM ECONOMIA

Teachers' grading patterns and student learning: evidence from São Paulo state public schools

Padrão de avaliação dos professores e aprendizado dos alunos: evidências das escolas estaduais de São Paulo

Fernando Amaral Carnaúba

Orientador: Prof. Dr. Marcos de Almeida Rangel

São Paulo - Brasil
2015

Prof. Dr. Marco Antonio Zago

Reitor da Universidade de São Paulo

Prof. Dr. Adalberto Américo Fischmann

Diretor da Faculdade de Economia, Administração e Contabilidade

Prof. Dr. Hélio Nogueira da Cruz

Chefe do Departamento de Economia

Prof. Dr. Márcio Issao Nakane

Coordenador do Programa de Pós-Graduação em Economia

FERNANDO AMARAL CARNAÚBA

Teachers' grading patterns and student learning: evidence from São Paulo state public schools

Padrão de avaliação dos professores e aprendizado dos alunos: evidências das escolas estaduais de São Paulo

Dissertação apresentada ao Departamento de Economia da Faculdade de Economia, Administração e Contabilidade da Universidade de São Paulo como requisito parcial para a obtenção do título de Mestre em Ciências.

Orientador: Prof. Dr. Marcos de Almeida Rangel

Versão Corrigida

(versão original disponível na Faculdade de Economia, Administração e Contabilidade)

São Paulo - Brasil

2015

FICHA CATALOGRÁFICA
Elaborada pela Seção de Processamento Técnico do SBD/FEA/USP

Carnaúba, Fernando Amaral
Teacher's grading patterns and student learning: evidence from São Paulo state public schools / Fernando Amaral Carnaúba. -- São Paulo, 2015.
55 p.

Dissertação (Mestrado) – Universidade de São Paulo, 2015.
Orientador: Marcos de Almeida Rangel.

1. Rendimento escolar 2. Teoria de resposta ao ítem 3. Avaliação da aprendizagem 4. Econometria I. Universidade de São Paulo. Faculdade de Economia, Administração e Contabilidade. II. Título.

CDD – 371.26

Agradecimentos

Agradeço à Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), à Fundação Instituto de Pesquisas Econômicas (FIPE), e à Fundação Lemann pelo apoio financeiro e institucional prestados durante a realização deste trabalho.

Aos professores Ricardo Madeira, Fernando Botelho, Eric Bettinger e Martin Carnoy, e ao colega Luan Paciencia pelas numerosas discussões que me ajudaram a compreender melhor as principais questões levantadas durante o trabalho.

Agradeço em especial ao professor Marcos Rangel, meu orientador neste trabalho, pela paciência, pela disponibilidade e pelos inúmeros conselhos oferecidos ao longo do percurso.

Aos amigos e à família, pelo pelo carinho e apoio sem tamanho. À Julia, por estar ao meu lado.

Resumo

Neste trabalho propomos um novo método para a medição do padrão de avaliação dos professores, fundamentado na Teoria de Resposta ao Item. Investigamos, com base no novo método, a relação entre o padrão de avaliação do professor e o aprendizado do aluno. Nós simulamos os potenciais resultados de uma política de aprovação de alunos baseada exclusivamente em um teste padronizado (Saresp), que implicaria em um padrão de avaliação único para cada série e disciplina, em substituição à política atual em que os professores são responsáveis pela definição sobre a aprovação de seus alunos. Estimamos os padrões de avaliação ótimos sob esta política, do ponto de vista da maximização do aprendizado dos alunos, e comparamos estes valores com os padrões de avaliação estimados para cada professor. Nossas estimativas indicam que os professores utilizam atualmente padrões de avaliação que são, em média, mais lenientes do que o padrão de avaliação ótimo estimado para a nova política.

Palavras-chaves: rendimento escolar, teoria de resposta ao item, avaliação da aprendizagem, econometria.

Abstract

We propose a new method for measuring teacher grading standards that is based on the Item Response Theory framework, and investigate the relationship between teacher grading standards and student learning in São Paulo State public schools in light of this new approach. We simulate a policy in which student achievement in a standardized examination (Saresp) is used as the passing grade criterion, setting a unique grading standard for each grade and subject that would substitute the current teacher-defined grading. We estimate the optimal standards that maximize student achievement under this policy, and compare them with the standards estimated for each individual teacher. Our estimates indicate that teachers currently apply standards that are, on average, more lenient than the optimal policy standards.

Key-words: school achievement, item response theory, standardized exams, econometrics.

Contents

1 Introduction

The impact of teachers' observable characteristics on the learning trajectories of students is a recurring theme in the economics of education literature. Several studies have uncovered empirical associations between teacher variables such as age, gender, experience, and the possession of professional degrees on student learning (Chetty, Friedman and Rockoff (2011), Darling-Hammond (2010), Hanushek and Rivkin (2006), Bettinger and Long (2010)). But a main obstacle faced by this literature remains - that many important characteristics of teachers are, in fact, unobservable to the researcher.

It has also been shown by the recent empirical literature in quasi-experimental settings that the provision of performance feedback to students influences achievement. Using data on high-school students in the Basque Country, Azmat and Iriberri (2010) find that the provision of feedback information leads to an average 5% increase in student scores. Similar conclusions are drawn by Bandiera, Larcinese and Rasul (2012), based on student records in a leading university in the United Kingdom. They estimate that the provision of feedback increases student performance by 13% on average. Moreover, the authors show how feedback works as an incentive mechanism that elicits heterogeneous responses across students. Their results indicate that while high-achieving students apply more effort when they receive feedback, low achieving students tend not to change effort levels.

The combined results of those two strands in the literature raise the point that an important unobserved characteristic of teachers is the method and rigor of evaluation used in the classroom, as well as the method of disclosure of achievement to students. While several choices involved in teaching are imposed as an external decision by the state-level authority and are fairly verifiable - such as curriculum and classes schedules - grading is mostly an unverifiable action and arguably constitutes the major incentive tool to be used in the classroom at the instructor's discretion.

In that direction, a small literature has started to shed light in the importance of grading patterns to student learning, focusing on whether higher grading standards

improve student outcomes (Haz, 2012, Figlio and Lucas, 2004, Betts and Grogger, 2003, and Betts, 1997). As described in more detail in the next section, this literature has focused on a univariate measure of teacher grading patterns based on the difference between mean achievement of students in the classroom and in external evaluations.

We propose a new method to quantify teachers' grading patterns that differs slightly from this literature. Our new strategy relies on an established solution to an analogous problem faced by the Test Theory literature: the measurement of scoring patterns of items in multiple choice tests. Test Theory has tackled the issue of item scoring characterization with Item Response Theory (IRT) models.

In light of this new approach, we analyze a rich database of public schools from the state of São Paulo in Brazil to investigate the impact of teacher grading patterns on student learning trajectories. We shed light on possible policy implications, including the simulation of a policy that centralizes student grading, by substituting a central standardized examination for the current teacher-based grading. Our contribution to the literature is twofold: (i) we are the first paper to analyze the relationship between student effort and teacher grading patterns for the Brazilian case, and (ii) we argue that through a more interpretable modelling of the relationship between student effort and teacher grading patterns, our new approach allows for more precise policy recommendations.

Nevertheless, several caveats should be noted and we recommend that our results be interpreted with caution. First, our analysis is restricted to the cognitive dimension of student abilities and focuses only on the maximization of mean test scores. Several other objectives should be taken into account in any policy decision such as noncognitive abilities, retention rates and distributional effects, to name a few. Second, we find evidence that our empirical approach may suffer from endogeneity problems which may lead to biased estimates. As a result, we suggest that our study be interpreted only as a starting point to further discussion about how to improve grading in Brazil and point out to some possible directions for future research.

The remainder of this paper is organized as follows. Chapter 2 reviews the previous

literature on grading patterns. Chapter 3 presents the databases used in this study. Chapter 4 presents a theoretical framework. Chapter 5 describes the empirical strategy. Chapter 6 presents the results. Chapter 7 presents counterfactual scenarios and discusses policy implications. Chapter 8 discusses the relationship between our results and the previous literature, and Chapter 9 concludes.

2 Previous literature on teacher grading patterns and student effort

The major obstacle to the investigation of the impact of grading patterns on student learning is that grading patterns are not directly observable to the researcher. It is necessary, thus, to recur to an indirect form of measurement. Albeit with some minor variations, the existing literature shares a common method for measuring grading patterns (Haz, 2012, Figlio and Lucas, 2004, Betts and Grogger, 2003, and Betts, 1997). Their main strategy is to compare the average scores of students in school with their scores in external standardized exams. With minor variations, in those studies the grading patterns are computed as the teacher fixed-effects coefficients in a regression of student standardized scores ($score$) on classroom marks ($class\_mark$)[1]. This empirical procedure leads to the following estimation equation:

$$score_i = \sum_{j=1}^{N_{teachers}} [\alpha_j teacher_{ij}] + \beta class\_mark_i + \epsilon_{ij} \tag{2.1}$$

Where i indexes students, j indexes teachers, $teacher_{ij}$ are teacher dummy variables and α_j are teacher grading pattern coefficients. In this definition, the higher the value of a teacher grading pattern, the more rigorous he is in assigning marks to his students. The intuition behind the estimation is that the teacher-fixed effect can be interpreted as a shift in students $score$, given his classroom mark. A large positive fixed effect for teacher j, thus indicates that his students are expected to present higher standardized scores for any given classroom mark.

To investigate the relationship between grading patterns and student effort, the existing literature also resorts to a common framework with minor variations. It consists of a linear regression of students score gains in standardized exams on teacher grading

[1] In order to avoid confusion, we use the term *classroom mark* or *mark* to refer to grades obtained by students in the school, as assigned by their teachers, and save the term *grade* for references to curriculum years (1st grade, 2nd grade, and so forth). Student achievement in standardized external exams is referred to as *score*.

standards, as described in the following equation:

$$\Delta score_{i,t} = \beta standard_{ij,t} + \varphi Z_{s,t} + \varepsilon_{ij,s,t} \tag{2.2}$$

Where i indexes students, j teachers, s schools and t time, $standard_{ij,t}$ is the grading standard faced by the student (i.e. coefficients α_j estimated in equation 1), and $Z_{s,t}$ is a vector of controls for school observable characteristics.

In the first known empirical study on the subject, Betts (1997) analyzed school-level grading patterns. Using data from the Longitudinal Study of American Youth (LSAY), he found higher grading standards to be associated with higher average achievement. Furthermore, he finds grading standards to have larger effects among high-achieving students.

Also measuring grading patterns at the school level, Betts and Grogger (2003) use data from the High School and Beyond survey to point out to likewise conclusions - higher grading patterns are associated with higher student scores, and larger effects are present among high achievers.

Figlio and Lucas (2004) were the first to measure grading standards at the teacher level. Analyzing data from a large school district in Florida, they also find higher grading standards to be associated with larger student score gains and larger effects among high-achieving students.

Haz (2012) estimates grading standards at the subject-school level, using a national database of 4th graders in Chile. As in Figlio and Lucas (2004), she finds higher grading standards to be positively associated with student score gains. In her study, though, larger gains are found among students at the bottom of the distribution of baseline scores. A main empirical advancement in Haz (2012) is that she corrects for the endogeneity in equation (2.2) by using past standards as an instrument to current standards.[2]

[2] Equation (2.2) is endogenous because the current *score* appears on both sides of the equation. On the left-hand side $score_{i,t}$ composes the measurement of student achievement gain, $\Delta score_{i,t}$. On the right-hand side, it takes part in the calculation of $standards_{ij,t}$. This endogeneity results on a positive bias on the estimation of β. Note that the effect of a random positive shock in $score_{i,t}$ is positive

for both $\Delta score_{i,t}$ and $standards_{ij,t}$. Conversely, the effect of a negative random shock in $score_{i,t}$ is negative for both $\Delta score_{i,t}$ and $standards_{ij,t}$. Measurement error in $score_{i,t}$ thus forces a positive association between $\Delta score_{i,t}$ and $standards_{ij,t}$, biasing β upwards.

3 Data

This research makes use of three databases: (i) administrative records of student achievement in school (marks assigned by teachers); (ii) a standardized and blindly corrected examination administered by the state-level authority (Saresp); and (iii) administrative records of teacher-classroom assignments[1]. The databases cover the years between 2008 and 2011.

Students' administrative records include marks assigned by teachers and attendance records for all students enrolled in the schools directly administered by São Paulo State Secretary of Education. Marks are subject-specific, assigned quarterly and summarized in a final mark that represents overall achievement through the year[2].

The standardized tests database consists of a statewide assessment, the São Paulo's Performance Evaluation System (*Sistema de Avaliação de Rendimento Escolar do Estado de São Paulo* – Saresp). The Saresp examination covers math, language and science[3], and is taken by students in the public system of education in grades 2 and 4 (elementary school), 6 and 8 (middle school), and 11 (high school). Student scores are calculated separately for each subject available using the Three Parameter Logistic Model of Item Response Theory, with the same scale of the nationwide exam Prova Brasil, which was set to have mean 250 and standard deviation 50 for the population in 8th grade that took this exam in in 1997. The scale used in Saresp is invariant from 4th to 11th grades, which allows scores to be comparable not only across years but also across most grades. The exam is also accompanied by a comprehensive socioeconomic survey. The scale is mapped into a set of learning standards for each grade and classified into proficiency levels, which are reproduced in Appendix A.

The teachers' administrative database consists of unique codes to identify assignments of teachers to classes. Teacher codes are consistent longitudinally, allowing for the

[1] All databases were shared with Prof. Marcos Rangel by São Paulo's Secretary of Education, under cooperation and confidentiality agreements.
[2] Only the final marks are used in this study.
[3] Only math and language are used in this study.

Table 1 – Descriptive statistics

Variable	Full Sample			Regression Sample			Difference	
	Mean	*sd*	*N*	*Mean*	*sd*	*N*	*Mean*	*p-value*
Panel A. 6th grade								
Math - Score	212.68	(39.08)	150,956	214.80	(38.38)	6,032	2.21	0.00
Math - Mark	5.91	(1.68)	151,567	5.70	(1.75)	6,032	-0.22	0.00
Language - Score	208.13	(42.33)	150,867	207.99	(41.94)	6,032	-0.16	0.78
Language - Mark	6.13	(1.66)	151,567	5.91	(1.71)	6,032	-0.22	0.00
Panel B. 8th grade								
Math - Score	246.46	(41.23)	146,515	246.69	(40.88)	11,275	0.26	0.52
Math - Mark	5.77	(1.75)	147,392	5.71	(1.82)	11,275	-0.06	0.00
Language - Score	231.77	(44.90)	146,623	233.29	(45.79)	11,275	1.64	0.00
Language - Mark	5.99	(1.70)	147,381	5.94	(1.75)	11,275	-0.06	0.00
Panel C. 11th grade								
Math - Score	270.22	(43.40)	99,810	270.31	(43.43)	1,121	0.09	0.94
Math - Mark	6.12	(1.55)	100,848	5.71	(1.68)	1,121	-0.41	0.00
Language - Score	269.47	(45.80)	100,167	266.76	(47.73)	1,121	-2.74	0.05
Language - Mark	6.30	(1.46)	100,830	5.83	(1.50)	1,121	-0.48	0.00

Note: All variables are presented in levels. *Score* refers to the Saresp examination, and is measured using IRT. *Mark* refers to final yearly classroom marks assigned by teachers, on a scale 0-10.

yearly identification of teacher-student assignment.

Summary statistics for our full dataset and our main sample of interest are shown in Table 1. The sample of interest is restricted to students for whom we have information on all variables necessary for our final estimations[4]. Even though several variables show a significant difference in means, the magnitudes of the differences are small in terms of standard deviations of the full sample. The largest relative differences are present in language and math marks, with respective sample differences of .33 and .27 standard deviations in 11th and .13 and .13 standard deviations in 6th grade. For all other variables, mean differences between the full and restricted samples are smaller than .05 standard deviations.

[4] Grade, score, lagged score, teacher and classroom mark for both math and language.

4 Theoretical framework

4.1 Measuring teacher grading patterns with IRT

We propose a new approach to quantifying teacher grading patterns that differs slightly from the framework used in the previous literature. Our approach draws from the test measurement theory, which has established Item Response Theory modelling as a standard method for measuring grading patterns of items. The core underlying assumption of IRT is that the probability of a student correctly answering a given item of an examination is a monotone and increasing function of his (latent) ability level. Each item is thus characterized by a particular function that maps latent ability levels of students into probabilities of correctly answering the item, called the Item Characteristic Curve (ICC). Therefore, the characterization an item's scoring pattern amounts to the complete specification of its ICC (Baker, 1992). For example, if an item k assigns a lower probability of correctly answering than item l for all possible levels of latent abilities, than item k is unequivocally more $difficult$ than item l. We transpose this idea to the measurement of teacher grading patterns – so we model the probability of scoring above a threshold in the classroom marks as a monotone function of student ability.

Thus, in order to quantify teacher grading patterns under the IRT framework, a first necessary step is to dichotomize marks, for example, into pass/fail or high/low achievement[1]. The highest category is then interpreted analogously to a "correct answer", or a "success" in a standard IRT model. As a result, the characterization of a teacher's grading pattern amounts to the complete specification of his ICC. A second step is to choose the functional form of the ICC. We adopt the Two-Parameter Logistic Model

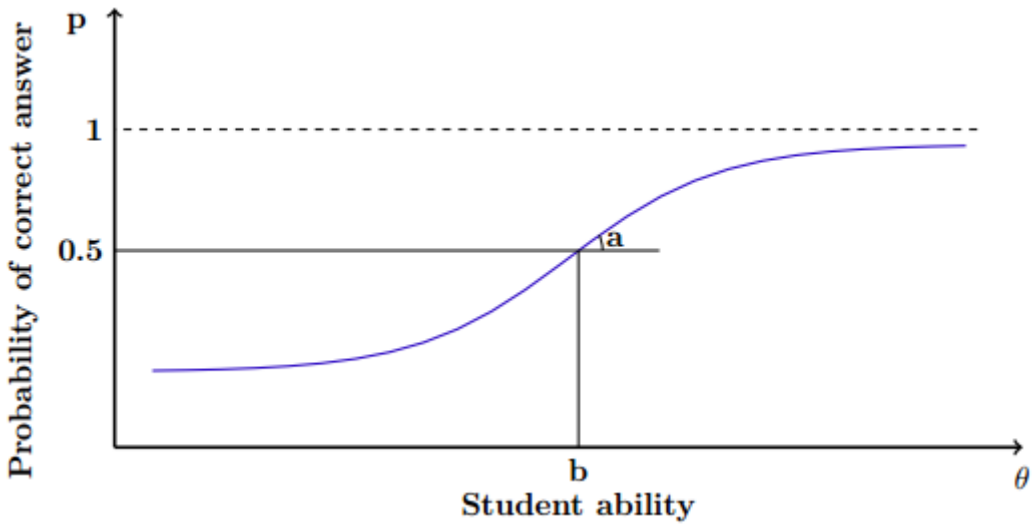
[1] Some IRT models allow for polytomous responses, but we focus on dichotomous models for the sake of simplicity.

(2PLM)[2].

$$P(success_i|\theta_i; a_j; b_j) = \frac{1}{1 + e^{-a_j(\theta_i - b_j)}} \tag{4.1}$$

As the name suggests, in this model the ICC of each teacher j is fully characterized by two parameters, which we describe using the conventional notation in the literature and the pass/fail dichotomization example. Parameter b, called $difficulty$ (or location) provides a measurement of the overall difficulty imposed by the teacher. More precisely, b is the latent ability level required for a student to have exactly 50% chance of receiving a passing grade from his teacher. Parameter a, called *slope*, defines the rate of transformation of latent abilities into probabilities of receiving a passing grade, and is directly proportional to the steepness of the ICC at point b. Figure 1 illustrates a general 2PLM characteristic curve.

Figure 1 – Item Characteristic Curve (ICC)



In standard IRT models it is usual to jointly estimate item parameters and student

[2] The Saresp examination uses tre Three Parameter Model (3PLM), which is the most common model used in multiple choice exams. The 3PLM includes a term called "guessing parameter" which is interpreted as the minimum probability of correctly answering the item. The inclusion of this term is anchored in the idea that even students with very low ability levels have a non-zero probability of assigning a correct answer by pure guess, since the number of possible answers is finite. We argue that the same reasoning does not apply for teacher grading patterns and thus opt for the 2PLM, which is equivalent to setting the "guessing parameter" to zero in the 3PLM.

abilities using a Maximum-Expectation algorithm (Schilling and Bock, 2005, Bock and Aitkin, 1981)[3]. But if reliable estimates of student abilities are available and only the item parameters are left to be estimated, a simple maximum likelihood estimation can be run separately for each item.

4.2 Student behavior

We assume students derive different utility levels from obtaining a *success* or $fail$ grade in the classroom, and that studying is costly[4].

$$v(e_i) = \begin{cases} u(success_i) - c(e_i), \text{ if } class\_mark_i \geq k \\ u(fail_i) - c(e_i), \text{ if } class\_mark_i < k \end{cases} \tag{4.2}$$

Where $u(success_i) - u(fail_i) > 0$, $e_i \geq 0$ and $c(e_i)$ is a continuous and twice differentiable function, with $c'(e_i) > 0$ and $c''(e_i) < 0$. Using an affine transformation to normalize $u(success_i) = 1$ and $u(fail_i) = 0$, and assuming students are risk-neutral, utility is given by[5]:

$$U(e_i) = P_j(success_{ij}|\theta_i) - c(e_i) \text{ s.t. } \theta_i = \theta_{0i} + e_i \tag{4.3}$$

Where $P_j(success_{ij}|\theta_i)$ is a teacher-specific function that maps effort and baseline scores (θ_i) into probabilities of obtaining *success*. Taking advantage of the IRT framework used to model teacher grading patterns, the probability of success is given by the 2PLM characteristic curve set by each teacher, represented in equation (4.4). We use a quadratic polynomial form to characterize the convexity of the cost function, as expressed in equation

[3] There are several other methods for the estimation of IRT models. See Wirth and Edwards (2007) for a comprehensive review.
[4] Our model draws, in part, from Haz (2012).
[5] The affine transformation is given by $U(e_i) = (v(e_i) - u(fail_i))/(u(success_i) - u(fail_i))$

(4.5).

$$P_j(success_{ij}|\theta_i) = \frac{1}{1+e^{-a_j(\theta_i - b_j)}} \tag{4.4}$$

$$c(e_i) = \frac{\gamma}{2}(e_i)^2 \tag{4.5}$$

Substituting equations (4.4) and (4.5) in (4.3) and framing the student decision on his choice of score θ leads to the following individual maximization problem:

$$\theta_1 = argmax_\theta \left[\frac{1}{1+e^{-a(\theta-b)}} - \frac{\gamma}{2}(\theta-\theta_0)^2\right] \text{ s.t. } \theta_i \geq \theta_{0i} \tag{4.6}$$

with first and second-order conditions:

$$\text{FOC: } \frac{ae^{-a(\theta-b)}}{(1+e^{-a(\theta-b)})^2} - \gamma(\theta-\theta_0) = 0 \tag{4.7}$$

$$\text{SOC: } \frac{2ae^{-2a(\theta-b)} - a^2e^{-a(\theta-b)}(1+e^{-a(\theta-b)})^2}{(1+e^{-a(\theta-b)})^4} - \gamma < 0 \tag{4.8}$$

Note that the student's optimal choice of effort is always positive, or, equivalently $\theta_i > \theta_{0i}$. This results from the fact that the marginal cost $c'(\theta)$ is zero at $\theta_i = \theta_{0i}$, while the marginal gain in utility from the increase in the probability of success, $P_j'(\theta)$ is positive at this (and any) point. Thus, there exists $\theta_1 > \theta_{0i}$ such that $U(\theta_1) > U(\theta_{0i})$.

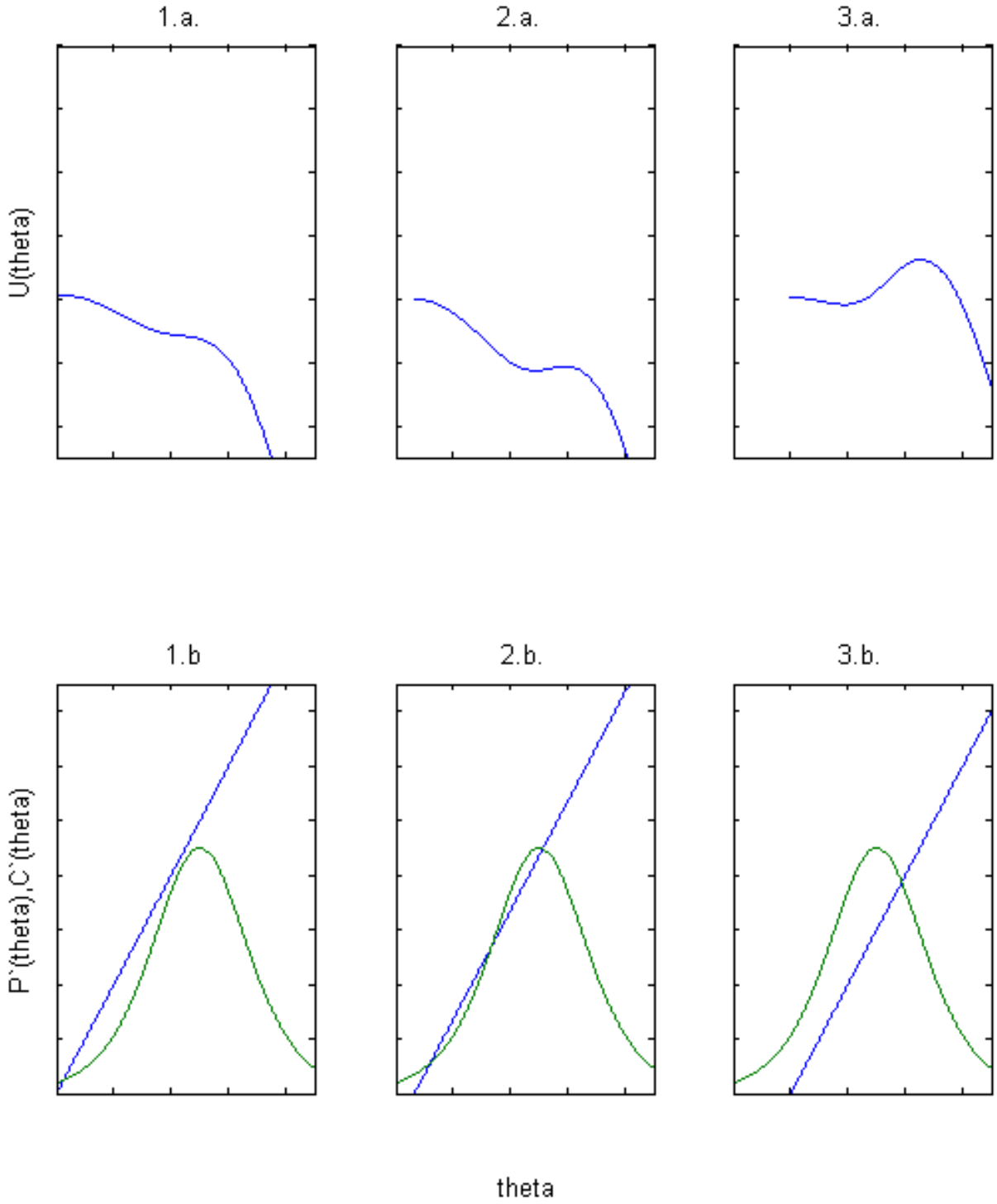
Figure 2 illustrates the choice of the optimal score gain (or analogously, effort) for students with different baseline scores. As exemplified in Panel 1, students with scores far below the difficulty parameter (b) will exert very low effort, because $P_j'(\theta)$ is very low for values of θ around the student initial level of ability (note that $lim_{\theta\to-\infty}P_j'(\theta) = 0)$). Those are the students that will start the year with a low probability of *success*, and will not try to increase it by much because "the bar is set too high" for them. Conversely, students

with scores far above the difficulty parameter (b) will also chose to exert very low effort, because $P_j'(\theta)$ is also very low at their baseline scores (note that $lim_{\theta \to +\infty} P_j'(\theta) = 0$). This is the case of students that already begin the year with a probability of *success* that is close to 1, letting almost no room for gains from exerting effort.

Students with baseline scores closer to the difficulty parameter will be the ones who will exert lager effort, since they will be able to benefit from the section where the payoff $P_j'(\theta)$ is the highest. Panel 3 illustrates this situation.

We show a more ambiguous case in Panel 2, in which there are two local maxima, with fairly different effort levels. Which of them will be the global maximum will depend on the parameters of the problem. What is most important to stress is that there exists a baseline score $\tilde{\theta_0}$ for which both local maxima will lead to the same utility, so that the student will be indifferent between exerting a low and a high level of effort. This is the case in which "the bar is barely reachable". Thus, there will be a discrete increase in the level of effort chosen by students with baseline scores $\theta_0 < \tilde{\theta_0}$ in comparison to students with baseline scores $\theta_0 > \tilde{\theta_0}$.

Figure 2 – Student Maximization Problem



5 Empirical strategy

5.1 Teacher grading patterns

As described in Section 4.1, our measure of grading patterns requires classroom marks to be dichotomized into two categories (success/fail). We estimate grading patterns for every possible threshold, starting at the passing mark (below/above or equal 5) and going up to the highest available mark (below/equal 10) and analyze which of them provides the best fit to the data. For each threshold k, the ICC of teacher j is thus given by the 2PLM:

$$P(class\_grade \geq k|saresp; a_{jk}, b_{jk}) = \frac{1}{1 + e^{-a_{jk}(saresp - b_{jk})}} \tag{5.1}$$

A key fact to understanding the estimation procedure adopted is that we use student latent ability values that were already estimated externally to the context of the study. Those latent ability values are the very students' scores provided in their Saresp examinations, which, as we have mentioned, are calculated using IRT. With known student abilities, estimation of teacher ICC parameters amounts to maximizing a simple likelihood function, in which the likelihood associated with each observation is given by the probability of occurrence of the student's outcome (success/fail), conditionally on his latent ability. Furthermore, the two parameters of each teacher will only appear in the individual likelihoods associated with his respective students, allowing for the maximization procedure to be run separately for each teacher (and threshold).

Thus, we estimate teacher grading parameters through the maximization of the following log-likelihood for each teacher j and threshold k:

$$(\hat{a_{jk}}, \hat{b_{jk}}) = argmax_{a,b} \sum_{i \in N_j} \{Y_i \frac{1}{1 + e^{-a_{jk}(saresp_i - b_{jk})}} + (1 - Y_i)(1 - \frac{1}{1 + e^{-a_{jk}(saresp_i - b_{jk})}})\} \tag{5.2}$$

where:

$$\begin{cases} Y_i = 1, \text{ if } class\_mark_i \geq k \\ Y_i = 0, \text{ if } class\_mark_i < k \end{cases}$$

This is numerically equivalent to running one logit regression for each teacher and threshold, in which the running variable is the latent ability of her students (Saresp scores), and the dependent variable is the dichotomous success/fail variable.

5.2 Student behavior

Since we have estimated teacher grading patterns (a_j and b_j) and observe both baseline scores ($saresp_{j,t-2}$) and current scores ($saresp_{j,t}$), the student cost coefficient γ is the only unknown parameter left to be estimated in order to fully characterize the student decision problem in equation (4.6). This parameter has a very specific interpretation: it is the rate of growth of the marginal cost of studying (measured in gains in the probability *success*), in terms of acquired ability (measured in the Saresp scale). Moreover, the underlying theoretical framework informs us that γ should be positive. We estimate γ using a non-linear least squares estimation framework.

Consider the student maximization problem in equation (4.6). Even though the problem admits no closed-form solution for θ_1, we can define a function (to be estimated numerically) that maps student input variables θ_0, a_{ij}, b_{ij} and the parameter γ into the chosen level of ability θ:

$$f(\theta_0, a_{ij}, b_{ij}, \gamma) = argmax_\theta \left[\frac{1}{1 + e^{-a_{ij}(\theta - b_{ij})}} - \frac{\gamma}{2}(\theta - \theta_0)^2 \right] \text{ s.t. } \theta \geq \theta_0 \tag{5.3}$$

The non-linear least squares estimator for γ can then be defined as:

$$\hat{\gamma} = argmin_{\gamma} \left[\sum_{i \in N} (f(\theta_0, a_{ij}, b_{ij}, \gamma) - \theta_1)^2 \right] \tag{5.4}$$

Substituting (5.3) in (5.4) and writing the equation in terms of our data, we arrive at the following NLS estimation:

$$\hat{\gamma} = argmin_{\gamma} \left\{ \sum_{i \in N} \{ argmax_{\theta} \left[\frac{1}{1 + e^{-a_{ij}(\theta - b_{ij})}} - \frac{\gamma}{2} (\theta - \theta_0)^2 \right] - saresp_{i,t} \}^2 \right\} \tag{5.5}$$

which we estimate numerically using brute-force algorithms for both the inner maximization of the student utility and the outer minimization of squared residuals[1].

[1] The decision of which range to use for the inner maximization is straightforward: we use the range assumed by Saresp scores in our full database. In the case of γ, even though there is no theoretical argument to define a closed interval in which it should be contained, empirically it becomes evident that for sufficiently large values of γ, student's change in optimal response tends to zero. That is, the cost of effort becomes so high that the student's optimal choice is numerically equivalent to zero, and variations cease to be captured in the inner student maximization.

6 Results

6.1 Teacher grading patterns

Tables 2 and 3 present descriptive statistics for parameters a and b estimated from equation (5.2). The fact that our standardized score is measured on a common scale for all grades in the data allows us to run a validity check of our measured grading standards: in theory, as students advance to further grades in school, teachers are expected to require increasing levels of knowledge from students in order to reward them with a given classroom mark. Taking the pass/fail dichotomization as an example (≥ 5), teachers are expected to require an increasing level of knowledge in order to reward students with a passing grade, which should be reflected as higher estimated b parameters for higher grades. We show in Table 3 that the estimated standards follow this pattern – that is, for every dichotomization (from ≥ 5 to ≥ 10), the median value of b across teachers increases as students move up to higher grades.

The only exception to this is the threshold 5 (passing grade) for 11th grade, which is lower than that of 8th grade for both math and language. This may be related to the fact that 11th grade is the last in secondary school. As such, passing 11th grade leads to the attainment of an important labor market signal which is a formal and informal reference in hiring. Secondary degree completion is required, for example, in the vast majority of civil servant positions. As a result, teachers may be reluctant to retain students.

When we compare the estimates of b with proficiency levels defined by Saresp (reproduced in Appendix A), a general pattern emerges that median coefficients for the passing grade (≥ 5) are always lower than the *below basic* reference score. This implies that there are students with ability level *below basic* who have more than 50% chance of receiving a passing grade. This is especially pronounced in 11th grade, in which the median b (166.1 for math and 144.2 for language) is more than two standard deviations below the *below basic* threshold (275 and 250, respectively).

As for the highest dichotomization of grades (threshold 10), the median b coefficient

is in general very close to the *advanced* threshold. This suggests that students who are considered advanced in Saresp tend to also be rewarded with the highest possible mark in their classrooms.

Another general pattern is that the slope parameter a increases for higher thresholds, in any given grade and subject. At first, this could indicate that teachers may interpret increases in the Saresp scale differently at different points of the scale. As we have pointed out, higher thresholds are associated with higher b parameters, so when we compare a coefficients across different thresholds we are actually comparing the rate of transformation of Saresp scores into classroom grades at different points of the scale. The pattern would then imply that teachers' scales are, in general, more granular for higher levels of ability. If this were the case, we would expect a to also be increasing in grades, because the b parameter is also increasing in grades for any given threshold. The data shows a different pattern, though, with a decreasing in higher grades. A second possible explanation – which we are not able to asses from the available data - is that for lower thresholds, teachers take more into account factors external to the cognitive ability level measured in Saresp such as class participation and group work.

6.2 Student behavior

We estimate equation (5.5) by NLS for each possible threshold, separately by grade and subject. Results are show in table 4. Consistently with the theoretical model, estimates of the cost parameter γ are positive for all thresholds and grades.

As in the case of parameter a, we observe a pattern of increasing values of γ for higher thresholds, for any given grade and subject. That raises the same possibility that have explored for a – that the scale of Saresp may be "compressed" for in the higher end of the measurement spectrum, in this case not by teachers but by students. That would mean that, in general, acquiring one ability point at the higher end of the distribution would be more costly than at the lower end of the distribution. If this were the case, we should expect γ to increase in higher grades, holding the threshold fixed. When we compare the

Table 2 – Estimates of teacher's grading parameters a (slope)

	Math				Language			
	Median	*Mean*	*sd*	*N*	*Median*	*Mean*	*sd*	*N*
Panel A. 6th grade								
Threshold								
≥ 5	0.034	0.119	0.470	57,599	0.035	0.200	0.699	48,946
≥ 6	0.035	0.042	0.026	83,169	0.034	0.041	0.027	74,403
≥ 7	0.039	0.069	0.204	83,096	0.036	0.051	0.092	74,504
≥ 8	0.047	0.335	1.022	76,552	0.039	0.243	0.803	69,314
≥ 9	0.057	0.895	1.945	55,898	0.045	0.679	1.614	49,431
≥ 10	0.074	1.265	2.213	24,096	0.051	1.073	2.031	19,911
Panel B. 8th grade								
Threshold								
≥ 5	0.028	0.050	0.148	60,215	0.030	0.133	0.519	54,284
≥ 6	0.028	0.032	0.019	80,733	0.028	0.033	0.020	72,432
≥ 7	0.031	0.038	0.028	80,743	0.030	0.037	0.027	72,464
≥ 8	0.038	0.256	0.814	74,254	0.035	0.173	0.603	67,085
≥ 9	0.046	0.702	1.576	55,400	0.043	0.640	1.530	47,592
≥ 10	0.061	1.063	1.849	25,853	0.054	0.929	1.756	18,787
Panel C. 11th grade								
Threshold								
≥ 5	0.022	0.035	0.060	21,823	0.022	0.046	0.159	16,859
≥ 6	0.020	0.022	0.013	44,237	0.019	0.023	0.015	37,615
≥ 7	0.022	0.026	0.016	45,092	0.022	0.026	0.017	39,818
≥ 8	0.026	0.084	0.305	41,306	0.028	0.080	0.316	35,665
≥ 9	0.035	0.449	1.116	29,558	0.037	0.459	1.258	23,628
≥ 10	0.048	0.868	1.561	13,585	0.044	0.777	1.667	8,707

Note: Estimates are run separately for each teacher, grade and threshold. Sample size N refers to the number of students.

values of γ across grades, though, we find that the parameter assumes its lowest values in 8th in general, not in 11th grade.

Table 3 – Estimates of teacher's grading parameters b (difficulty)

	Math				**Language**			
	Median	*Mean*	*sd*	*N*	*Median*	*Mean*	*sd*	*N*
Panel A. 6th grade								
Threshold								
≥ 5	152.9	140.1	57.1	57,599	140.6	127.8	58.0	48,946
≥ 6	210.6	208.9	28.8	83,169	192.9	191.1	30.3	74,403
≥ 7	240.5	241.1	26.8	83,096	228.7	229.2	29.1	74,504
≥ 8	264.1	268.0	30.7	76,552	259.2	263.7	35.7	69,314
≥ 9	282.1	287.7	34.2	55,898	283.4	291.1	43.3	49,431
≥ 10	293.7	298.5	32.3	24,096	294.7	304.7	46.3	19,911
Panel B. 8th grade								
Threshold								
≥ 5	177.4	159.5	70.0	60,215	154.1	137.2	66.8	54,284
≥ 6	251.3	250.3	35.3	80,733	220.4	218.4	35.9	72,432
≥ 7	283.9	285.8	34.7	80,743	261.1	262.9	35.9	72,464
≥ 8	308.1	314.1	40.6	74,254	293.5	299.3	42.3	67,085
≥ 9	325.4	332.1	39.7	55,400	314.9	323.7	45.2	47,592
≥ 10	339.9	347.5	44.2	25,853	326.1	334.2	41.1	18,787
Panel C. 11th grade								
Threshold								
≥ 5	166.1	120.9	128.8	21,823	144.2	114.6	106.8	16,859
≥ 6	261.1	256.1	66.6	44,237	238.0	229.7	68.2	37,615
≥ 7	308.8	312.1	56.3	45,092	297.5	296.4	55.0	39,818
≥ 8	343.6	352.2	57.6	41,306	338.1	344.0	57.0	35,665
≥ 9	361.6	378.3	65.5	29,558	356.7	373.0	63.6	23,628
≥ 10	372.6	381.6	53.6	13,585	367.6	379.2	55.0	8,707

Note: Estimates are run separately for each teacher, grade and threshold. Sample size N refers to the number of students.

Table 4 – Non-linear Least Squares estimates of students' gamma parameters (cost)

Threshold	**$\geq$ 5**	**$\geq$ 6**	**$\geq$ 7**	**$\geq$ 8**	**$\geq$ 9**	**=10**
Panel A. 6th grade						
Gamma - Math	0.00034	0.00113	0.32768	1.31072	0.0051	0.1800
R2	0.22	0.07	0.06	0.08	0.15	0.22
Gamma - Language	0.00027	0.00061	0.00128	0.02048	0.0410	0.5700
R2	0.35	0.21	0.19	0.18	0.25	0.25
N	6,032	9,992	9,410	8,295	4,333	880
Panel B. 8th grade						
Gamma - Math	0.00014	0.00028	0.00043	0.00075	0.0013	0.0200
R2	0.20	-0.17	-0.31	-0.32	-0.30	-0.19
Gamma - Language	0.00024	0.00043	0.00064	0.00184	0.0029	0.0200
R2	0.42	0.34	0.30	0.27	0.29	0.31
N	11,275	18,806	18,979	16,835	10,108	2,744
Panel C. 11th grade						
Gamma - Math	0.00016	0.00048	0.00171	0.00236	0.0051	0.9100
R2	0.19	0.04	-0.02	0.05	0.09	0.19
Gamma - Language	0.00012	0.00026	0.00032	0.00073	0.0015	1.3107
R2	0.34	0.23	0.13	0.09	0.09	0.19
N	1,121	3,617	3,843	3,127	1,663	434

Note: Gamma parameters were estimated separately for each threshold, grade and subject. Samples are the same across subjects, given threshold and grade.

7 Counterfactual analysis

7.1 Optimal grading standards for some reference classrooms

Consider the decision problem of a teacher who wants to choose his grading parameters in order to maximize the mean achievement of his N_j students in the Saresp examination. Students have a common cost parameter γ and individual baseline abilities θ_{0i}, which are all observable to the teacher:

$$(a_j^*, b_j^*) = argmax_{a,b} \sum_{i \in N_j} \theta_{1j} \text{ s.t. } \frac{ae^{-a(\theta_{1j}-b)}}{(1+e^{-a(\theta_{1j}-b)})^2} = \gamma(\theta_{1i} - \theta_{0i}) \tag{7.1}$$

We study the solution to (7.1) for simulated classroom settings that aim to represent general patterns of 6th grade language classrooms. Results are shown in Table 5 and illustrated in Figure 3. We set the classroom size to 30 and test 9 combinations of mean and variance of baseline scores, using the parameter γ estimated for 6th grade language. The classroom mean increases from row 1 to rows 2 and 3 - we add 50 and 100 points[1] on each subsequent row, and set row 2 to match the mean of 6th graders in language. As for variances, we start with a constant distribution (column 1), on which we perform two mean preserving spreads, the first with the imposition of a uniform distribution of range 150 (column 2), and the second with the imposition of a uniform distribution of range 300 (column 3)[2].

As expected, increases in the mean baseline scores lead the optimal b parameter to increase by the same amount, while the optimal a remains unchanged. Increases in variance, on the other hand, imply a lower optimal b. This results from the fact that students optimal choice of effort is asymmetrical in relation to the distance between their baseline score and b. Students with baseline scores below b exert more effort than students with scores above b, for any given distance. Thus, when we perform a mean preserving

[1] The standard deviation of 6th grade language scores in our full sample is 42.33.

[2] This implies standard deviations of 46 and 91, respectively, which are respectively approximate to 1.1 and 2.2 standard deviations of 6th grade language scores in our full sample.

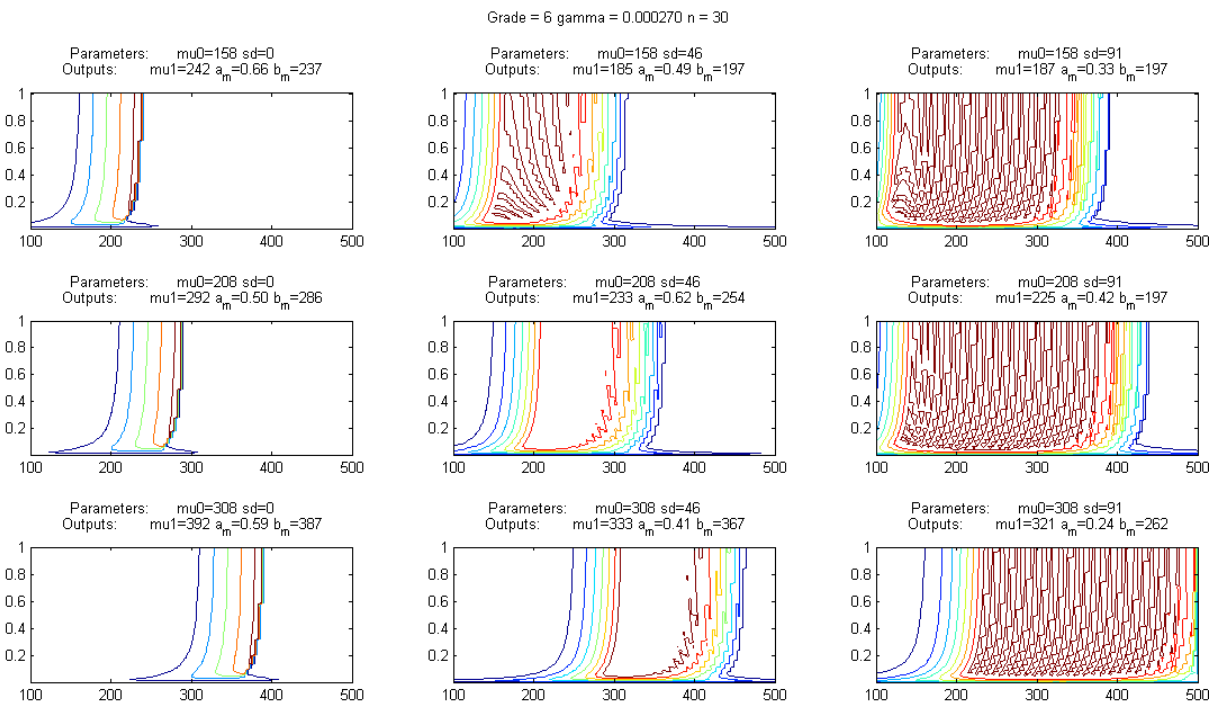
Table 5 – Mean Student Achievement Under Different Class Setting an Grading Parameters

	$\mathbf{mean}(\theta_0) = \mathbf{158}$			$\mathbf{mean}(\theta_0) = \mathbf{208}$			$\mathbf{mean}(\theta_0) = \mathbf{308}$		
	$mean(\theta_1)$	a	b	$mean(\theta_1)$	a	b	$mean(\theta_1)$	a	b
$\mathbf{sd}(\theta_0) = \mathbf{0}$	242	0.66	237	292	0.5	286	392	0.59	387
$\mathbf{sd}(\theta_0) = \mathbf{46}$	185	0.49	197	233	0.62	254	333	0.41	367
$\mathbf{sd}(\theta_0) = \mathbf{91}$	187	0.33	197	225	0.42	197	321	0.24	262

spread in the classroom distribution of baseline scores, the loss in effort from students that shift further from b is larger for students below b than for students above b. As a result, it is optimal for the teacher to lower the chosen difficulty parameter.

Our analysis of the optimal grading pattern for different classroom settings leads thus to two conclusions. First, holding classroom variance of baselines scores fixed, grading standards should be set higher for higher average baseline scores. Second, holding classroom average baseline scores fixed, grading standards should be set lower in classrooms with higher baseline score variance.

Figure 3 – Mean Student Achievement Under Different Class Setting an Grading Parameters



7.2 Optimal unique grading standards and policy discussion

Consider the decision problem of a policy maker who wants to choose a unique grading pattern to be used by all teachers that maximizes the mean achievement of all N students in Saresp. Students share a common cost parameter γ and individual baseline abilities θ_{0i}, which are all observable to the policy maker:

$$(a_j^*, b_j^*) = argmax_{a,b} \sum_{j \in J} \sum_{i \in N_j} \theta_{1j} \text{ s.t. } \frac{ae^{-a(\theta_{1j}-b)}}{(1+e^{-a(\theta_{1j}-b)})^2} = \gamma(\theta_{1i} - \theta_{0i}) \tag{7.2}$$

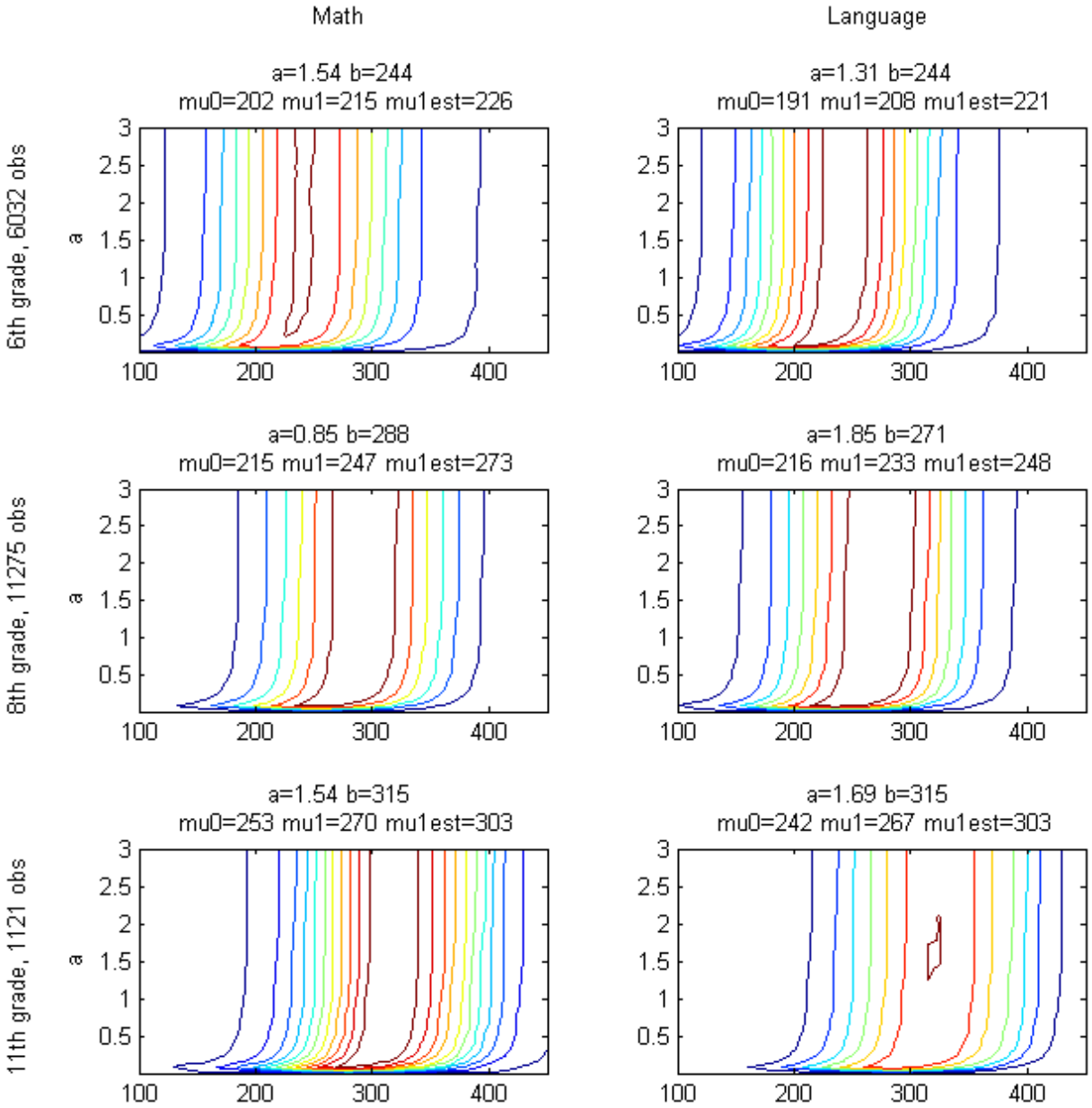
It is important to emphasize that this problem characterizes a policy that substitutes Saresp for the current grading pattern. That is, it simulates a situation in which Saresp would be used as the passing grade criterion, replacing students classroom marks. Alternatively, it can be interpreted as the upper-bound estimate of the expected results of merely disclosing Saresp scores to students.

Interestingly, there is no unambiguous theoretical prediction as to whether the optimal unique standard policy should lead to higher average achievement than the current teacher autonomy policy. On one hand, setting a unique standard could guarantee that the (unique) optimal standard is implemented in every classroom. This could lead to an improvement in effort of several students who may currently be assigned to a teacher whose standards are sub-optimal. On the other hand, setting a unique standard is a very strong restriction in comparison to the teacher-defined current grading policy, in which several different standards can be used to effectively adapt for the optimal setting at each classroom. This loss in degrees of freedom could lead to an overall decrease in student effort if teachers are currently applying grading standards that do not depart much from the optimum.

We solve the problem numerically[3] for the passing grade threshold, using our estimates of γ for each grade and subject, and our data on student baseline levels of ability. Results are shown in Table 6 and illustrated in Figure 4.

[3] We use a brute-force algorithm with a range that is larger than range of estimated teachers' a and b parameters.

Figure 4 – Mean Student Achievement Under Unique Standards Policy



Our estimates of the impact of the new policy in mean student achievement are positive for all grades and subjects. The implied increases in mean scores ranges are fairly large, ranging from .28 (11 points in 6th grade math) to .99 (43 points in 11th grade language) standard deviations of the respective grade and subject full sample scores.

All estimated a^* coefficients are fairly higher than the range of median estimates for teacher parameters across grades. In fact, the estimated coefficients are high enough to virtually lead to a perfect separation of students below the thresholds, who would have probability close zero of receiving a passing grade, and students above the thresholds, who

Table 6 – Mean Student Achievement Under Unique Standards Policy

	a	**b**	**mean(θ_0)**	**mean(θ_1)**	**mean($\hat{\theta}_0$)**
Math					
6th grade	1.54	244	202	215	226
8th grade	0.85	288	215	247	273
11th grade	1.54	315	253	270	303
Language					
6th grade	1.31	244	191	208	221
8th grade	1.85	271	216	233	248
11th grade	1.69	315	242	267	303

would have probability of passing close to one.

Estimates for b^* are significantly higher than the estimates of teacher grading patterns - for all grades and subjects, b^* is more than one standard deviation above the correspondent median teachers. In all cases, our estimates of passing-grade b^* are closer to teachers median values of thresholds 7 and 8 – in fact, for all grades and subjects b^* is between the correspondent median b for thresholds 7 and 8 of estimated for teachers.

As discussed in Section 6.1, a main motivation for teachers to implement passing grade patterns with b coefficients that are sub-optimal may be the reduction of retention rates. If this is the case, we should expect the observed grading patterns to be further from the optimum in grades 8th and 11th than in 6th, because São Paulo State public schools were under a policy of automatic approval for students in this grade. We have also pointed out that stakes involved in graduation in 11th are arguably the highest across the three grades we analyze, suggesting that the b should be the furthest from b^* in this grade. Our results match this description. Math median teacher estimates are found to be further from b^* as we move from 6th to 8th and 11th grades (91.1, 110.6 and 148.9 points, respectively), as well as language estimates (103.4 , 116.9 and 170.8, respectively).

This pattern raises the point that retention rates are an important dimension related to grading, and should be taken into account in actual policy-making. We suggest this as an important venue to be explored in future studies.

The policy simulation exercise makes it clear that the focus of our analysis is

restricted to average student achievement on cognitive math and language standardized tests, with the assumption that this is the only dimension taken into account in the policy maker's objective function. Teacher's current grading patterns may in fact take several other cognitive and noncognitive dimensions into account, providing a richer set of incentives that is not captured by our model. Future research may shed light into this relationship and help analyse other dimensions of grading from a policy perspective.

8 Relation to the previous literature

We now replicate the empirical strategy used in the previous literature with our data. Teacher grading patterns are estimated by taking the difference in mean student achievement in standardized test scores and in the classroom, which is one of the variations of equation (2.1) used in the literature. Imposing $\beta = 1$ in equation (2.1), estimates of grading patterns simplify to:

$$score_i = \sum_{j=1}^{N_{teachers}} [\alpha_j teacher_{ij}] + class\_mark_i + \varepsilon_{ij} \tag{8.1}$$

We then analyze the relationship between grading standards with the same linear regression specification used in previous studies. Two different dependent variables are tested: current scores (equation (8.2)) and gains in scores (equation (8.3)).

$$saresp_{j,t} = \alpha + \beta standards_{k,t} + \varepsilon_{j,t} \tag{8.2}$$

$$\Delta saresp_{j,t,t-2} = \alpha + \beta standards_{k,t} + \varepsilon_{j,t} \tag{8.3}$$

We also replicate the identification strategy of Haz (2012), using lagged standards as instruments for current standards. Estimates are presented in Table 7. Columns (1) and (2) show results for Math for equations (8.2) and (8.3), respectively, while columns (3) and (4) introduce lagged standards as instruments for current standards. Columns (5-8) mirror columns (1-4) for language.

The estimated coefficients that do not control for both sorting (student fixed-effects) and endogeneity (instrumental variables) are universally positive, indicating that higher rigor in grading is associated with increases in student scores. As we introduce controls for student fixed-effects and instrumental variables, though, the effect is reduced, indicating that sorting and endogeneity are important and should be controlled for. In the

Table 7 – Regressions of Scores on Teacher Grading Patterns

	Math				**Language**			
	OLS		*IV*		*OLS*		*IV*	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A. 6th grade								
Standards	0.458***	0.152***	0.316***	0.052	0.407***	0.0558*	0.309***	-0.0781
	(0.0316)	(0.0271)	(0.0514)	(0.0409)	(0.0392)	(0.0314)	(0.0519)	(0.0446)
N	6,032	6,032	4,685	4,685	6,032	6,032	4,474	4,474
R-squared	0.040	0.007	0.038	0.003	0.029	0.001	0.031	0.00
Panel B. 8th grade								
Standards	0.527***	0.151***	0.429***	0.000	0.457***	0.0974***	0.348***	-0.003
	(0.0275)	(0.0211)	(0.0407)	(0.0320)	(0.0294)	(0.0212)	(0.0437)	(0.0293)
N	11,275	11,275	9,235	9,235	11,275	11,275	8,750	8,750
R-squared	0.045	0.005	0.043	0.000	0.032	0.002	0.028	0.00
Panel C. 11th grade								
Standards	0.396***	0.0498	0.309***	-0.052	0.450***	0.159***	0.295***	0.152**
	(0.0635)	(0.0545)	(0.0750)	(0.0706)	(0.0594)	(0.0453)	(0.0804)	(0.0692)
N	1,121	1,121	1,024	1,024	1,121	1,121	1,015	1,015
R-squared	0.033	0.001	0.029	0.00	0.062	0.011	0.044	0.012
Student FE		yes		yes		yes		yes

Note: Standard-errors are clustered at teacher level. *** significant at 1%.

specification that controls for both sorting and endogeneity estimates are inconclusive at the 5% significance level.

In our policy simulation we are able to control for sorting, since we use student score gains to account for student fixed-effects. On the other hand, we are not able to control for endogeneity issues. This raises an important caveat to our study. Our results may be subject to the same positive bias found in our replication of the models used by the previous literature, and thus should be interpreted with caution.

9 Conclusion

In this study we have proposed the use a of new method for measuring teacher grading patterns based on the IRT framework, which is widely used in test theory but to our best knowledge has not yet been applied to teacher grading. We argue that this method allows for a more interpretable measurement of grading patterns.

Comparing our estimates of teacher grading patterns with standards published by São Paulo state-level education authority, we find that the median teacher tends to reward students who are graded at the highest category of the Saresp scale (*excellent*) with the highest possible classroom mark (10). This matching in scales on the higher end of the ability scope is reassuring, especially when we take into account that teachers do not observe student's Saresp scores. This matching is not replicated, however, in the lower end of the ability scope. We find that the median teacher has a high probability of assigning passing grade scores for students below the lowest category in the Saresp scale (*below basic*).

We simulate a policy in which the Saresp scores is used as the passing grade criterion, setting unique grading standards for each grade and subject that would substitute the current teacher-defined grading. The estimated effects of the policy are fairly high, suggesting an increase in mean scores that ranges from .28 to .99 standard deviations across the grades and subjects analyzed. Our optimal difficulty parameters (b) are significantly above the parameters estimated for the median teacher, which implies that retention rates would be increased under the new policy.

Together, our policy simulation and the comparison of estimated grading patterns with the official Saresp scale suggest that teachers value a distributional dimension in their choices of grading parameters, in particular with respect to retention, at the expense of the maximization of average scores. We do not intend to judge the adequacy of this choice, and restrict our analysis to the measurement of only one side of the tradeoff - the impact of grading on average scores. We suggest the investigation of the impact of grading patterns on retention as an important venue to be explored in future studies.

We emphasize a number of caveats and recommend that our results be interpreted with caution. First, our analysis is methodologically restricted to the cognitive dimension of student abilities and focuses only on the maximization of mean test scores, thus ignoring several other objectives that should be taken into account in any policy decision. Second, our empirical approach may suffer from endogeneity problems that may lead to biased estimates. As as a result, we suggest that our study be interpreted only as a starting point for further discussion about how to improve grading in Brazil.

Bibliography

ANDRADE, D. F. de; TAVARES, H. R.; VALLE, R. da C. Teoria da resposta ao item: conceitos e aplicações. *ABE, Sao Paulo*, 2000.

ARROW, K. J. Models of job discrimination. *Racial discrimination in economic life*, Lexington, Mass.: Lexington Books, v. 83, 1972.

AZMAT, G.; IRIBERRI, N. The importance of relative performance feedback information: Evidence from a natural experiment using high school students. *Journal of Public Economics*, Elsevier, v. 94, n. 7, p. 435–452, 2010.

BAKER, F. B. Item response theory: Parameter estimation techniques. Marcel Dekker New York, 1992.

BAKER, F. B.; KIM, S. *Item response theory: Parameter estimation techniques*. [S.l.]: CRC Press, 2004.

BANDIERA, O.; LARCINESE, V.; RASUL, I. Blissful ignorance? a natural experiment on the effect of feedback on students' performance. *Labour Economics*, Elsevier, v. 34, p. 13–25, 2015.

BECKER, G. S. *The economics of discrimination*. [S.l.]: University of Chicago press, 2010.

BETTINGER, E.; LONG, B. T. *Do college instructors matter? The effects of adjuncts and graduate assistants on students' interests and success*. [S.l.], 2004.

BETTINGER, E.; LONG, B. T. Do faculty serve as role models? the impact of instructor gender on female students. *American Economic Review*, JSTOR, p. 152–157, 2005.

BETTINGER, E.; LONG, B. T. Does cheaper mean better? the impact of using adjunct instructors on student outcomes. *The Review of Economics and Statistics*, MIT Press, v. 92, n. 3, p. 598–613, 2010.

BETTS, J. R. Do grading standards affect the incentive to learn? UCSD Economics Discussion Paper 97-22, 1997.

BETTS, J. R.; GROGGER, J. The impact of grading standards on student achievement, educational attainment, and entry-level earnings. *Economics of Education Review*, Elsevier, v. 22, n. 4, p. 343–352, 2003.

BOCK, R. D.; AITKIN, M. Marginal maximum likelihood estimation of item parameters: Application of an em algorithm. *Psychometrika*, Springer, v. 46, n. 4, p. 443–459, 1981.

BOTELHO, F.; MADEIRA, R.; RANGEL, M. Discrimination goes to school? racial differences in performance assessments by teachers. *Unpublished Manuscript. Department of Economics, University of Sao Paulo*, 2010.

BOTELHO, F.; MADEIRA, R.; RANGEL, M. *Discrimination Goes to School*. [S.l.], 2014.

CHETTY, R.; FRIEDMAN, J. N.; ROCKOFF, J. E. The long-term impacts of teachers: Teacher value-added and student outcomes in adulthood. nber working paper no. 17699. *National Bureau of Economic Research*, ERIC, 2011.

DARLING-HAMMOND, L. Teacher quality and student achievement. *Education policy analysis archives*, v. 8, p. 1, 2000.

FIGLIO, D. N.; LUCAS, M. E. Do high grading standards affect student performance? *Journal of Public Economics*, Elsevier, v. 88, n. 9, p. 1815–1834, 2004.

GROSSMAN, P. et al. *Measure for measure: The relationship between measures of instructional practice in middle school English language arts and teachers' value-added scores*. [S.l.], 2010.

HANNA, R. N.; LINDEN, L. L. Discrimination in grading. *American Economic Journal: Economic Policy*, American Economic Association, v. 4, n. 4, p. 146–168, 2012.

HANUSHEK, E. A.; RIVKIN, S. G. Teacher quality. *Handbook of the Economics of Education*, Elsevier, v. 2, p. 1051–1078, 2006.

HAZ, V. A. P. Essays on the economics of education. 2012.

KANE, T. J.; STAIGER, D. O. *Estimating teacher impacts on student achievement: An experimental evaluation*. [S.l.], 2008.

KLEIN, R. Utilização da teoria de resposta ao item no sistema nacional de avaliação da educação básica (saeb). *Revista Meta: Avaliação*, v. 1, n. 2, p. 125–140, 2009.

KLUEGER, A. N.; DENISI, A. Effects of feedback intervention on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin*, v. 119, n. 2, p. 254–284, 1996.

PHELPS, E. S. The statistical theory of racism and sexism. *The American Economic Review*, JSTOR, p. 659–661, 1972.

RANGEL, M. Is parental love colorblind? human capital accumulation within mixed families. *The Review of Black Political Economy*, Springer, v. 42, n. 1-2, p. 57–86, 2015.

SCHILLING, S.; BOCK, R. D. High-dimensional maximum marginal likelihood item factor analysis by adaptive quadrature. *psychometrika*, Springer, v. 70, n. 3, p. 533–555, 2005.

SOARES, T. M. Influência do professor e do ambiente em sala de aula sobre a proficiência alcançada pelos alunos avaliados no simave-2002. *Estudos em Avaliação Educacional*, n. 28, p. 103–124, 2013.

SPERLING, M.; FREEDMAN, S. W. Research on writing. *Handbook of Research on Teaching*, v. 4, p. 370–389, 2001.

THORNDIKE, E. L. Human learning. The Century Co, 1931.

WIRTH, R. J.; EDWARDS, M. C. Item factor analysis: current approaches and future directions. *Psychological methods*, American Psychological Association, v. 12, n. 1, p. 58, 2007.

Appendix A

Table 8 – Proficiency level classification on Saresp scale

Panel A. Proficiency Level Thresholds

Proficiency Level	**Language**			**Math**		
	6th *grade*	8th *grade*	11th *grade*	6th *grade*	8th *grade*	11th *grade*
Below basic	<175	<200	<250	<200	<225	<275
Basic	175 to <225	200 to <275	250 to <300	200 to <250	225 to <300	275 to <350
Adequate	225 to <275	275 to <325	300 to <375	250 to <300	300 to <350	350 to <400
Advanced	$\geq$ 275	$\geq$ 325	$\geq$ 375	$\geq$ 300	$\geq$ 350	$\geq$ 400

Panel B. Description of Proficiency Levels

Classification	*Proficiency Level*	*Description*
Insufficient	Below basic	Students demonstrate insuficient proficiency on content, competencies and abilities desirable for their current grade/year.
Sufficient	Basic	Students demonstrate minimal proficiency of content, competencies and abilities, but have the structures required to interact with the curriculum of the subsequent grade/year.
	Adequate	Students demonstrate full proficiency of content, competencies and abilities desirable for their current grade/year.
Advanced	Advanced	Students demonstrate proficiency on content, competencies and abilities above the required level for their current grade/year.