Conveying Value Via Categories

Paula Onuchic ⓡ Debraj Ray[†]

October 2019

Abstract. A sender sells an object of unknown quality to a receiver who pays his expected value for it. Sender and receiver might hold different priors over object quality. The sender commits to a monotonic categorization of quality in order to maximize her expected value from the sale. We characterize the sender's optimal categorization for any pair of sender and receiver priors and propose an algorithm which, under a mild restriction on those priors, generates finitely many pooling and separating intervals in the optimal categorization. If priors are smooth, the solution involves alternate zones of pooling and separation. We provide conditions under which full pooling is optimal, as well as conditions under which full separation is optimal. In an application, we study the design of an educational grading scheme by a profit-maximizing school, which seeks to signal student qualities and simultaneously incentivize students to learn. These incentive constraints can be embedded into a reduced-form school objective as a distortion of the school's prior over student qualities, and can thus be rewritten as a particular categorization problem with distinct sender and receiver priors.

1. Introduction

A sender is about to come into possession of an object of unknown quality. Prior to knowing that quality, she commits to a *categorization*. That is, she partitions the set of qualities into subsets — some possibly singletons — and verifiably commits to reveal the element in which the quality belongs. The categories must be monotone. For instance, she can place objects of quality between a_0 and a_1 into one category, and objects of quality above a_1 in another. She cannot, however, lump together qualities below a_1 or above a_2, where $a_2 > a_1$. Monotonicity is an assumption, but in our main application, we will derive this property from a more primitive setting.

A receiver buys the object, obtains value equal to quality (realized after purchase), and pays the sender his expected value for the object, conditional on the sender's category announcement. The sender seeks to maximize the expected value paid by the receiver.

The sender and receiver use distinct distributions to evaluate the expectation of quality. An obvious interpretation is that they hold different priors, and the main part of the paper

[†]Onuchic: New York University, p.onuchic@nyu.edu; Ray: New York University and University of Warwick, debraj.ray@nyu.edu. We are grateful to Heski Bar-Isaac, Navin Kartik, Laurent Mathevet and Ennio Stacchetti for useful comments. Ray acknowledges funding under National Science Foundation grant SES-1851758. "ⓡ" indicates author names are in random order.

can be read with this interpretation. But, as we shall see, there are other situations that generate the same reduced form. For instance, the sender might be working on behalf of a subgroup that already possesses some private information about quality, and so want to maximize expected payoff with respect to their interim belief. Or there could be separate incentive constraints that serve to effectively distort the measure that the sender maximizes expected value with respect to, *even* though the priors are common. Indeed, it is even possible that the resulting“measure” isn't a distribution at all. Temporarily postponing this last somewhat cryptic observation, notice that a difference in priors is what makes the categorization problem nontrivial — if sender and receiver have the same prior, nothing is gained by committing to any categorization.

Our main result (Theorems 1 and 2) fully describes the profit-maximizing categorization for any pair of initial distributions over object quality held by sender and receiver. Moreover, under a mild joint restriction on the shapes of the two prior distributions, we describe an algorithmic procedure which mechanically generates the pooling and separating zones identified in Theorem 2 (Proposition 1).

When both the priors have differentiable cumulative functions, Proposition 2 shows that the solution involves alternate zones of pooling and separation. (Without differentiability, it is possible to have adjacent pooling intervals with no separating zone in between.) If we view the pools as discrete grades assigned to a continuum of potential qualities, then the separating regions that lie between two grades may be interpreted a deliberate choice to render precise descriptions of qualities that lie close to the grade boundaries.

Proposition 3 provides necessary and sufficient conditions for full pooling to occur; that is, for there to be no quality revelation at all. Conversely, Proposition 4 provides necessary and sufficient conditions for full separation to occur.

There are several applications of our categorization exercise; after all, intermediaries often provide consumers with information about the quality of purchased goods or services. To name just a few examples, financial rating agencies classify assets according to their riskiness, certifying companies emit eco-friendly labels, bond issues are rated by certification agencies such as Moody's, the Department of Health provides restaurants with sanitary inspection grades, and schools grade students according to their academic achievements. We choose to focus on a particular application — educational grades — because this emphasizes some interesting features of the core model.

First, educational grades play a dual role: they signal student ability, but they also incentivize learning, which might have intrinsic value over and above ability. In this sense, the problem for the school is a mixture of mechanism *and* information design. Of course, this feature is not restricted to education alone. For instance, Moody's rating structure allows lenders to learn about the inherent risks of bond issuers, while incentivizing them to take less risky actions. One might wonder how these concerns are nested by our admittedly spartan model, which has no moral hazard.

Second, an education setting highlights the constraints naturally faced by designers that lead to monotone categorization. In our application, non-monotone categorizations would not satisfy the incentive constraint — higher-ability students could always put in less effort and get the same grade as their lower-ability counterparts.

Third, any education design model must make explicit what schools (or senders) are trying to maximize. In our case it is tuition revenue. But tuition revenue is related to the surplus a school can extract from its students, without violating their participation constraints. Here, the receiver (an employer) pays the *student*, not the school, and the student pays the school. However, what a student is willing to pay the school depends on her own prior regarding ability. If different students have different priors, the "lowest belief type" among admitted students must pin down the equilibrium tuition rate. It is this belief that effectively becomes the prior of the school, even though its own beliefs — and those of the receiver — might coincide with the population distribution of abilities.

Using Propositions 5–7, Theorem 3 shows how we can fold all these considerations — including moral hazard — into a special case of our baseline model, with an appropriately induced sender prior. Sender-receiver priors become different (even if they are the same in the original setting). Indeed, the induced sender prior may not be a cdf at all.

This nesting generates some insight into the school design problem. Proposition 8 extends Proposition 3 to show when a school might choose to keep student quality unrevealed by a deliberate policy of grade inflation. Conditions for pooling are more propitious when learning has little intrinsic value (an observation related to Lizzeri 1999), when the effort costs of schooling are accounted for in the optimal tuition rate (students — or more likely their parents — may incompletely account for these), and when the lowest belief type has more information about her own ability. Proposition 9 explores a special case in which the school uses a mixture of a single separation zone and a single pooling zone, each active at different points of the ability and learning distribution. The school chooses a threshold ability type below which all students are pooled in a single grade, exert no effort and have no learning and above which all types are revealed and learning is increasing in ability type. We show that the threshold type is decreasing in the optimism of the lowest belief type is; that is, the school optimally chooses to reveal more information to the market by expanding the separating region. However, the threshold type is non-monotone in the extent to which the market values pure learning. In this last case, schools could therefore respond by revealing more, or revealing less.

2. Related Literature

Our paper contributes to the large and growing literature on "information design," stemming from Rayo and Segal (2010) and Kamenica and Gentzkow (2011). In one sense, our problem is a special case of the Bayesian persuasion model, because our payoffs are linear in state and action. So the value of a categorization depends only on the distribution of posterior means it induces. On the other hand, our major departure is that we allow for sender and receiver to hold distinct priors, which is certainly not part of the original setup. This is most closely related to Alonso and Camara (2016), who extend the concavification method to the case of heterogeneous priors. As has been widely noted in the literature, the concavification method alone is typically not sufficient to characterize the optimal signal, especially when the space and action space is large as in our model. We provide an algorithmic procedure that generates the optimal monotone signaling structure for any pair of priors held by the sender and the receiver. Galperti (2019) and Kartik, Lee and Suen (2019) also study environments with heterogeneous priors across agents. Both these papers study problems different from ours, but some results in Kartik, Lee and Suen (2019) have parallels in our paper that have to do with the special question of when full pooling or separation of qualities is optimal.[1]

Kolotilin (2018) and Dworczak and Martini (2019) also study optimal signaling structures. They provide verification methods for the optimality of some candidate structure. Sender and receiver have a common prior. The main tension arises from differences in the way quality is valued: the payoff to the sender is some nonlinear function of expected quality. This model is distinct from our setting with heterogeneous priors and linear payoff. Neither can be mapped into the other. Finally, our problem is also distinct in that the sender is restricted to choose a monotone information structure, while Kolotilin (2018) and Dworczak and Martini (2019) deal with the optimality of a structure in the class of *all* signaling structures, monotone or not. Our restriction is *not* a special case of a model in which non-monotone structures are allowed — the search for a monotone structure that is optimal in the class of all monotone structures is a distinct exercise, and one that is often justified by the application at hand, such as our education example. For more discussion, see Section 3.1. We also note that one of the results in Dworczak and Martini (2019) provides conditions for the optimality of monotone structures, though again for the case of a common prior. Kolotilin and Li (2019) provide some characterization results for monotone persuasion for the sender payoff being in a a class of nonlinear functions of the expected quality and sender and receiver sharing the same prior.

Since Lizzeri (1999), a growing literature studies the incentives to conceal information by grading agents coarsely. For example, Skreta and Veldekamp (2009) study ratings inflation in financial markets and Ostrovsky and Schwarz (2010), Popov and Bernhardt (2013) and Boleslavsky and Cotton (2015), among others, study grade inflation in the educational context. In most of this literature, agents are passive and payoff directly

[1] If sender and receiver have likelihood-ratio ordered priors, Kartik, Lee and Suen (2019) find that full pooling is optimal for the sender. Conversely, if the sender's prior dominates the receiver's, full revelation is optimal. We show that if the receiver prior first-order stochastically dominates the sender's prior, this is enough for full pooling, but the converse is not true — if the sender's prior first-order stochastically dominates that of the receiver, it is not enough for full revelation.

depends on the types of agents who are being graded. That is also true of our baseline model. Yet our application studies a case in which receiver payoffs depend on student ability and learning, where the latter is endogenously determined. Student quality is shaped not just by ability but by the exertion of costly learning effort. Grade categorization plays the dual role of signaling ability and incentivizing student effort. This situation can be usefully mapped into our baseline, meaning that to some degree, applications with endogenous quality can be studied as a distortion of the sender's prior.

In our consideration of moral hazard, we are most closely related to Boleslavsky and Kim (2018), who also look at an environment in which signaling structures motivate costly effort. They extend the concavification method in Kamenica and Gentzkow (2011) to this environment. In our application, agents exert effort after drawing their ability and the incentive constraint implies that signaling structures must be monotone. This is not the case in Boleslavsky and Kim (2018), where costly effort improves the distribution of types that will be drawn from. As mentioned, it is not always easy to characterize the optimal signaling structure using the concavification method when the state and action spaces are large. In fact, most of the characterizations provided in Boleslavsky and Kim (2018) relate to the case of a binary state space or binary actions. Our procedure allows us to characterize the optimal structure that incentivizes agents and signals their abilities to the receivers in the case of a continuum of agent types and a continuum or receiver actions. Rodina and Farragout (2016) also study the problem of a principal who wants to improve an agent's investment in productivity when the only instrument at hand is an information disclosure policy. Their environment is different from ours both in how the agent's effort decision is set up and in the grading schemes that the authors allow for. In their environment, they find that in a wide variety of circumstances the optimal disclosure policy has a threshold form, i.e., is a monotonic categorization.

Moldovanu, Sela and Shi (2007) study the problem of a principal who wants to incentivize agents who are playing a contest for status to exert effort by designing "status categories". In the context of a contest for status, the principal is only designing how coarsely to order agents from best to worst, rather than concealing or revealing their actual qualities. They find that coarser schemes, such as putting agents in two categories, are optimal when the distribution of qualities is sufficiently concave. Dubey and Geanakoplos (2010) also study the design of categories to incentivize agents in games of status, allowing for both cardinal and ordinal comparisons. They too find that often agents are best motivated by coarse categories. Our abstract setting is free of moral hazard, though our application studies both adverse selection and moral hazard. In this latter case, we find that often the best way to motivate agents is by providing a mix of coarse and revealing categories. In the special case of a receiver with a uniform prior, coarser grading is related to a more concave sender prior over the distribution of types.

Doval and Skreta (2018) develop a toolbox for extending the concavification method to the case where the sender is subject to a certain class of incentive constraints and briefly review the literature where these tools apply. Lipnowski, Mathevet and Wei (2019) study a problem where the sender is constrained to the receiver's decisions to pay attention to the signals he is sent. In their model, the sender has the same objective as the receiver, but does not take into account attention costs incurred by the receiver. Similarly, in our application, the school does not fully incorporate the student's effort costs of learning. Guo and Shmaya (2019) study the problem of a sender who sends a message to a receiver with private information about his type in order to motivate him to take one of two actions. They show that an incentive compatible mechanism takes the form of nested intervals: each receiver type accepts on an interval of states and a more optimistic type's interval contains a less optimistic type's interval. They then develop tools to solve for the optimal signaling structure within this category of structures with nested intervals. In our paper, incentive compatibility similarly implies a restriction on the class of signaling structures available – in our case, to monotone structures.

After a first draft of our paper was written, we came across Rayo (2013), who studies the problem of a monopolist selling a status good to a population of heterogeneous agents, where agents of higher types are willing to pay more for the same good. This is closely related to our application. The monopolist offers a menu of goods at different prices and agents value not the inherent quality of a good, but the association to agents who also consume that same good. In that sense, an agent is willing to pay more for goods which are also consumed by agents of higher types. The monopolist's goal is to maximize his expected profit by designing a menu which pools some agent type intervals and separates others. As in our application, the agents incentive constraints restrict the monopolist to choosing monotone signaling structures. Rayo shows that the incentive constraints in his problem can be restated as a distortion of the monopolist's objective and that a solution to that problem consists of pooling in the *maximal intervals* where pooling is optimal and separating everywhere else.

3. MODEL

A sender is about to come into possession of an object of heterogeneous quality a, distributed according to a continuous distribution S with compact support $[\underline{a}, \bar{a}]$. A receiver buys the object from the sender and obtains value equal to its quality. The receiver believes that quality is distributed according to a continuous cdf R, strictly increasing on $[\underline{a}, \bar{a}]$ with $R(\underline{a}) = 0$ and $R(\bar{a}) = 1$. We assume that the receiver stands ready to pay the sender his expected value for the object, where expectations are computed using R and any information that the sender might have chosen to reveal.

One interpretation is that sender and receiver hold distinct priors. There are others, including the possibility that these priors are reduced forms of an extended model. In fact, our main application relies on the possibility that S may not be a cdf at all, but is instead some continuous function of bounded variation satisfying the endpoint conditions $S(\underline{a})$

finite and $S(\bar{a}) = 1$. While it is convenient to read this section by viewing S as a cdf, none of our results or formal arguments rely on such a presumption.

Prior to knowing the quality of the object, the sender commits to a monotonic quality categorization. That is, she creates (possibly singleton) *intervals* of qualities. For instance, she can create two categories — objects with quality between $\underline{a}$ and a_1 and objects with quality between a_1 and $\bar{a}$. She cannot, however, lump together qualities below a_1 or above a_2, where $a_2 > a_1$, into one category. The sender chooses these categories so as to maximize her revenue from the sale of the objects.

This problem can be equivalently written in the following way. Let $\mathcal{A}_R$ be the collection of all nondecreasing, right-continuous functions $A : [\underline{a}, \bar{a}] \to [\underline{a}, \bar{a}]$ such that

$$A(x) = \mathbb{E}_R \left[y | A(y) = A(x) \right]. \tag{1}$$

It is easy to see that this set is nonempty; e,g., $A(x) = \mathbb{E}_R(a)$ for all x lies in it. The sender picks a *categorization* $A \in \mathcal{A}_R$ to

$$\text{maximize} \int_{\underline{a}}^{\bar{a}} A(x) dS(x). \tag{2}$$

By way of interpretation, note first that any $A \in \mathcal{A}_R$, being nondecreasing, has (at most) countably many intervals on which it is strictly increasing and (at most) countably many intervals on which it is constant. Let's view these intervals as "exhaustive" in the sense that no strictly increasing interval is followed by a strictly increasing interval (it's all one interval) and a constant interval can only be followed by another constant interval if the constant values are different. By right-continuity,[2] these intervals are closed on the left and open on the right, except when there is an interval ending at the highest quality $\bar{a}$, and notation for intervals $[t, t')$ will be taken to mean just that.

3.1. **Remarks on Monotonicity.** First, a monotonicity restriction may be relevant in many situations. We may simply want to rule out some categories such as "this is either a junk bond or a AAA bond, but not anything in between," as being unrealistic. Second, in cases where there is an associated moral hazard problem, monotonicity emerges as the outcome of a broader mechanism design problem in which incentive constraints need to be respected. It becomes the expression of single-crossing. In Section 4, we describe such an application in detail.

Finally, it should be appreciated that the monotonicity restriction does not turn our model into a special case of a model that permits non-monotonic categories. To appreciate the distinction, suppose for a moment that all categories can be used. Then we can adapt the technique of Kamenica-Gentzkow to directly create a concavification argument

[2]The imposition of right continuity (left-continuity would done just as well) rules out isolated points for $A(x)$. Given the assumed continuity of R and S this does not matter at all.

that solves the categorization problem.[3] After all, the choice of arbitrary categories allows posteriors to be convexified to any required degree, so that any non-concavities in the seller's value function (viewed as a function of these posteriors) can be removed. These convexifications, achieved through belief-splitting, are no longer available when categories are monotone, necessitating a different approach.

3.2. **Categorization and Sender Value.** The categories implicit in any $A \in \mathcal{A}_R$ consists of all the constant or "pooling" intervals, and *every element* of every increasing or "separating" interval. Once told the categorization and the particular category in which the object lies, the conditional value of the object — in the receiver's eyes — is given by the expectation of a in that category element, under R. This is precisely equation (1). The sender chooses a categorization to maximize her expected payoff. As an example, suppose R is uniform on $[0,1]$ and S is uniform on a symmetric interval of length 2ϵ around $\{3/4\}$.[4] If the sender fully pools, then the associated $A(x)$ has constant value $1/2$ on $[0,1]$, since the expectation is calculated under R with no additional information. The sender then integrates over A with respect to S, and hence the sender's value under this policy is $1/2$, which is lower than the expectation of x under S. She could achieve the latter by committing to reveal quality; i.e., by choosing $A(x) = x$.

The sender can do still better, though. If she pools all qualities in the two intervals $[0, 3/4-\epsilon)$ and $[3/4-\epsilon, 1]$, then $A(x) = 3/8 - \epsilon/2$ if x lies in the former interval and $A(x) = 7/8 - \epsilon/2$ if $x \geq 3/4 - \epsilon$. The sender then integrates over A with respect to S, obtaining $7/8 - \epsilon/2$.

For any $A \in \mathcal{A}_R$, consider its pooling and separating intervals. Each such interval is of the form $[t, t')$ where t is the left edge and t' the right edge, with the right parenthesis closed if $t' = \bar{a}$. Define the *weighting* Ψ associated with categorization $A \in \mathcal{A}_R$ by:

$$
(3) \quad \Psi(x, A) = \begin{cases} S(x) \text{ if } x \text{ is in a separating interval;} \\ S(t) + [R(x) - R(t)] \left[\dfrac{S(t') - S(t)}{R(t') - R(t)} \right] \text{ if } x \text{ is in a pooling interval } [t, t'). \end{cases}
$$

Because R and S are continuous and R is strictly increasing, Ψ is well-defined and continuous. Moreover, $\Psi(\underline{a}, A) = S(\underline{a})$ and $\Psi(\bar{a}, A) = 1$. If S is a cdf, so is Ψ, though formally this is not needed. In either case, because S is a function of bounded variation, so is Ψ, and the integral in Observation 1 below is well defined.

Observation 1. *The value to the sender,* $\int_{\underline{a}}^{\bar{a}} A(x) dS(x)$*, equals* $\int_{\underline{a}}^{\bar{a}} x d\Psi(x, A)$.

[3]In fact, Alonso and Camara (2016) do adapt the Kamenica-Gentzkow concavification argument to the case of sender and receiver with heterogeneous priors.

[4]This ensures S is continuous. For the exposition, it might help to imagine that S is degenerate at $3/4$.

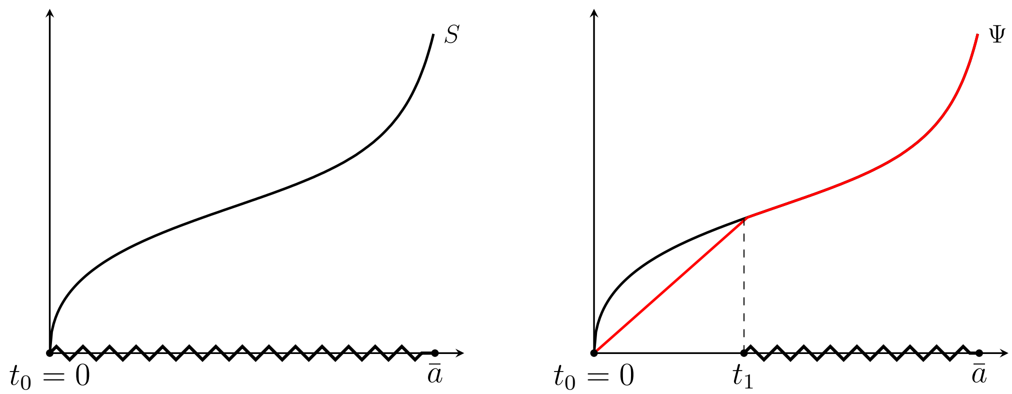


FIGURE 1. R is uniform on $[0, \bar{a}]$ and S is as pictured. The jagged lines indicate separating regions. In the left panel, under full separation, Ψ is equal to S. In the right panel, pooling between t_0 and t_1 and separating elsewhere, Ψ is given by the distribution in red. In the right panel, $t_1 \succ t_0$.

This observation states that the sender's value under any categorization A is equivalently found by integrating x over $[\underline{a}, \bar{a}]$ under the weighting $\Psi(\cdot, A)$, which in separating regions "follows" the sender's prior S and in pooling regions "follows" the receiver's prior R. Figure 1 shows $\Psi(\cdot, A)$ for two different categorizations A. See Section 5 for the proof of this, and other omitted proofs in the main text.

3.3. **Optimal Categorization.** For any continuous function. $H : [0, 1] \to \mathbb{R}_+$, define its *lower convex envelope* by

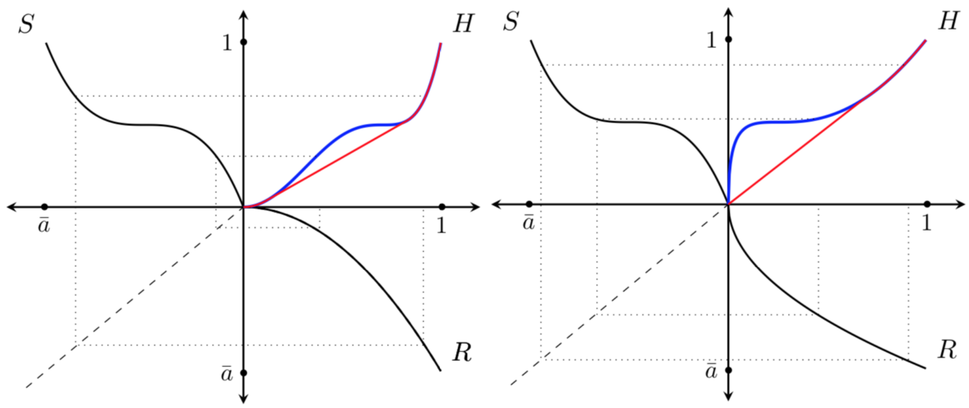
$$\breve{H}(x) \equiv \min\{y | (x, y) \in \text{Co}(\text{Graph}(H))\},$$

This chalks out — uniquely — the largest convex function we can place below H. In what follows, we study the particular function $H = S \circ R^{-1}$. Figure 2 constructs this function graphically. S (shown horizontally flipped) has a "reverse-logistic" shape. In panel A, R is concave (shown vertically flipped) and in panel B, R is convex (again vertically flipped). In each panel, the function $H = S \circ R^{-1}$ is derived, and displayed in the north-east quadrant. Its lower convex envelope is also displayed in each case.

To see how this function comes into play, first write any weighting function Ψ associated with a categorization A in *percentile weighting* form; that is, define for any $z \in [0, 1]$,

$$\Phi(z, A) \equiv \Psi(R^{-1}(z), A),$$

We can write this in a format parallel to (3) by defining a *percentile separating interval* of A as any interval $[w, w')$ such that $[R^{-1}(w), R^{-1}(w'))$ is separating, and a *percentile pooling interval* of A as any interval $[w, w')$ such that $[R^{-1}(w), R^{-1}(w'))$ is pooling.



(A) Concave Receiver Prior (B) Convex Receiver Prior

FIGURE 2. The construction of $H = S \circ R^{-1}$ and its lower convex envelope. In panel A, R is concave. In panel B, R is convex. In both cases, S has a reverse-logistic shape. H is shown in blue and its lower convex envelope in red.

(Remember to close the right parenthesis when $w' = 1$.) Then

$$\Phi(z, A) = \begin{cases} H(z) \text{ if } z \text{ is in a percentile separating interval;} \\ H(w) + (z - w)\left[\dfrac{H(w') - H(w)}{w' - w}\right] \text{ if } z \text{ is in some percentile pooling } [w, w'). \end{cases} \tag{4}$$

In particular, the percentile weighting associated with a categorization $A \in \mathcal{A}_R$ is equal to H in some (percentile separating) regions and is a straight line connecting $(w, H(w))$ and $(w', H(w'))$ in other (percentile pooling) regions of the form $[w, w')$. This means that $\text{Graph}(\Phi(\cdot, A)) \subset \text{Co}(\text{Graph}(H))$, which immediately implies:

Observation 2. *For any $A \in \mathcal{A}_R$ and $z \in [0, 1]$, $\Phi(z, A) \geqslant \breve{H}(z)$.*

Next, notice that $\breve{H}$ has zones where it coincides with locally convex segments of H (not necessarily all of them, though), and other intervals where it is simply a straight line connecting two points of the form $(z, H(z))$ and $(z', H(z'))$. We can thus fashion a categorization, the associated percentile weighting of which matches $\breve{H}$, by pooling the intervals where $\breve{H}$ is a straight line and separating everywhere else.

To see how this categorization behaves, return to Figure 2. If R is concave (panel A), this tends to compress the concave segment of the derived function H, and elongate the convex segment. The pooling interval of our categorization is therefore "small." This is reasonable: concavity of R signals pessimism on the part of the receiver about quality, so it is better that the sender separates qualities to a greater degree (relative to uniform R). In fact, Panel A shows two distinct zones of separation. In panel B, R is convex

— the receiver is relatively optimistic. This accentuates the concave segment of H and induces greater pooling, which is intuitive in the face of that optimism.

Observation 3. *There exists $A^* \in \mathcal{A}_R$ such that $\Phi(z, A^*) = \breve{H}(z)$ for all $z \in [0, 1]$.*

Observations 2 and 3 together imply the existence of a distinguished categorization A^* such that for every categorization $A \in \mathcal{A}_R$, and for every $z \in [0, 1]$, $\Phi(z, A) \geqslant \Phi(z, A^*)$. By a change of variables to recover the original weighting functions, we must conclude that for every $A \in \mathcal{A}_R$,

$$\Psi(x, A) \geqslant \Psi(x, A^*) \text{ for every } x \in [\underline{a}, \bar{a}]. \tag{5}$$

Recall that $\Psi(\underline{a}, A) = S(\underline{a})$ and $\Psi(\bar{a}, A) = 1$ irrespective of the categorization A. So the inequality (5), coupled with Observation 1, yields[5]

Theorem 1. *A categorization A^* with the property that $\Phi(z, A^*) = \breve{H}(z)$ for all $z \in [0, 1]$, where $H = S \circ R^{-1}$, is a solution to the sender's problem.*

Notice that finding the lower convex envelope of $S \circ R^{-1}$ is equivalent to finding the upper concave envelope of the inverse $R \circ S^{-1}$, and so this characterization would be equivalent to the *concavification* of this inverse, if S were invertible. However, to accommodate our main application, we cannot (and do not need to) make this assumption.

3.4. **A Characterization of A^*.** Our objective is to fully describe the sender's optimal categorization and its associated pooling and separating intervals. To that end, for any a and a', say that a' is a *successor* of a (or $a' \succ a$) if $a' > a$ and

$$\frac{S(x) - S(a)}{R(x) - R(a)} \geq \frac{S(a') - S(a)}{R(a') - R(a)} \tag{6}$$

for all $x \in (a, a']$. To understand this definition, suppose $a' \succ a$. Then conditional on placing all weight in $[a, a']$, R is "more optimistic" than S: it puts more weight on higher quality levels. So the sender might profit from treating $[a, a']$ as a pooling interval. For instance, in Figure 1, R is the uniform distribution on $[0, 1]$ and S is as pictured. The left panel has full revelation, which yields the expected value of a under S. In the right panel, there is pooling between t_0 and its successor t_1, and separation everywhere else. The resulting value is the expectation under the distribution in red, which follows R in the pooling region, and S thereafter. Since the red distribution is a first-order improvement over S, the expected value under this scheme is strictly higher. Hence, we can view intervals $[a, a']$, with $a' \succ a$, as those over which it is profitable to pool.Indeed, Panel B suggests that we can make further gains by moving t_1 to the right towards even larger successors of t_0. That motivates the following definition: $[a, a')$ is a *maximal interval* if

[5]Using integration by parts, we get $\int_{\underline{a}}^{\bar{a}} x d\Psi(x, A) = -(1 - S(\underline{a}))\underline{a} + \int_{\underline{a}}^{\bar{a}} (1 - \Psi(x, A))dx$, so that (5) implies that the sender value is weakly higher under A^* than under any $A \in \mathcal{A}_R$.

$a' \succ a$ and there is no pair $\{b, b'\} \neq \{a, a'\}$ such that $b \leqslant a$ and $b' \geq a'$, and $b' \succ b$. The following observation guarantees that maximality is a well-defined concept, and — crucially — that maximal intervals must be disjoint.

Observation 4. (i) *If a has a successor, it has a largest successor.*

(ii) *If $a_k \downarrow a$ and each a_k has a successor, then so does a.*

(iii) *Suppose that $a < b < a' < b'$, with $a' \succ a$ and $b' \succ b$. Then $b' \succ a$. In particular, no two maximal intervals can have a nonempty intersection.*

We can now state:

Theorem 2. *Choose $A^* \in \mathcal{A}_R$ so that it pools in all maximal intervals and separates in all other intervals. Then A^* is well-defined, and solves the sender's problem.*

Theorem 2 has two steps. First, A^* is well-defined. This requires Observation 4, particularly part (iii), which guarantees that no two maximal intervals overlap. The second step is that the collection of maximal intervals chalk out precisely the lower convex envelope of the composite function $H = S \circ R^{-1}$. That is, the envelope is made out of straight-line segments joining the edges $(w, H(w))$ and $(w', H(w'))$, where $[w, w')$ is maximal, and by following $H(z)$ otherwise. The proof is completed by appealing to Theorem 1.

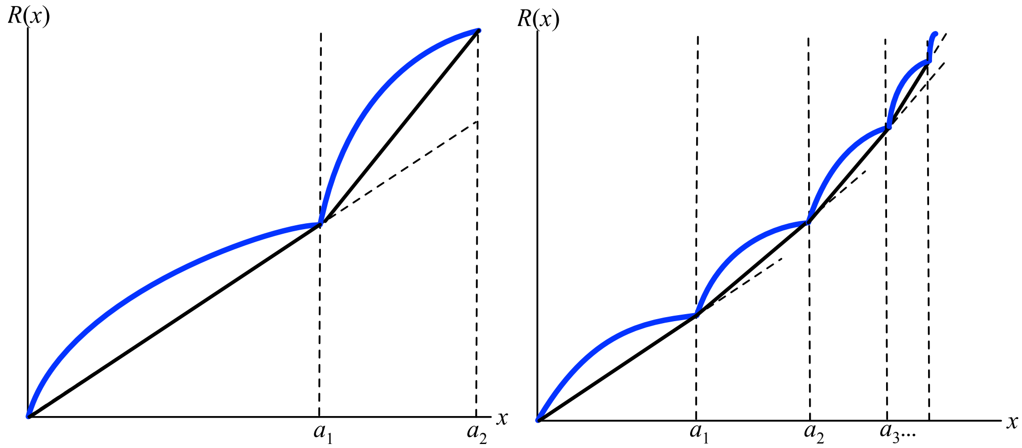
3.5. **A Procedure to "Compute" A^*.** The categorization A^* can be described by an algorithmic procedure if we assume:

[I] There are finitely many maximal intervals in $[\underline{a}, \bar{a}]$.

Despite the absence of a precise genericity statement, it is clear that Condition I will fail only for singular configurations of R and S. In Figure 3, R is uniform (not shown) and S is depicted by the "scalloped" blue line. In panel A, it is easy to see, using the definition in (6), that a_1 is the largest successor of $\underline{a}$, and by the same argument, that a_2 is the largest successor of a_1. But the situation could repeat itself, and could do so indefinitely, as depicted in Panel B. In this case Condition I fails. Conditional on R being uniform, this imparts a delicate asymptotic shape to S which would vanish on "small perturbations" of either function. This is why we view the Condition as "generic."

Under Condition I, the following procedure generates A^*, the verification of which is an easy consequence of Observation 4 and Theorem 2.

Proposition 1. *Begin at $a_0 = \underline{a}$. If a_0 has a successor, then set a_1 equal to the* largest *such successor (well-defined by part (i) of Observation 4), and designate $[a_0, a_1)$ to be a pooling interval. If a_0 has no successor, then a_1 is set to the first $a > a_0$ that* does *have a successor, or failing that, to $\bar{a}$. (This is guaranteed by part (ii) of Observation 4.) In that case, designate $[a_0, a_1)$ to be separating. If $a_1 = \bar{a}$, close the open right parentheses in the intervals above, and end the procedure.*



(A) Adjacent Pooling (B) Adjacent Pooling, Indefinitely

FIGURE 3. R is uniform (not shown) and S is the blue curve. Panel A clarifies, using (6), why there are two successive pooling intervals. Panel B shows that this could conceivably recur indefinitely, provided S has the "appropriate" shape; Condition I rules this out.

Otherwise, repeat the same process starting from a_1 to find $a_2, a_s \ldots$ Continue until $[\underline{a}, \bar{a}]$ is fully covered. Under Condition I, the procedure must end in finitely many steps.

3.6. **Separation Amidst Pooling.** When the cumulative functions S and R are differentiable, the optimal solution acquires an additional property:

Proposition 2. *Suppose R and S are differentiable, with strictly positive densities on $(\underline{a}, \bar{a})$. Then between two pooling intervals there is always a zone of separation.*

That is, if we view the pools as discrete grades assigned to a continuum of potential qualities, it is as if any cross-pool boundary has more precise annotations about product qualities. That is, optimal categorization involves a mix of coarseness and revelation. One way to interpret these alternating zones is that coarse categories are created, but within those coarse categories, the best or worst quality objects are revealed. A plus or minus annotation in student grades would loosely conform to this description, as also letters of recommendation in which students receiving the high end of a particular grade are often pointed out for having done so. In financial ratings, often rating agencies report "outlooks" on assets' ratings, signaling that either an asset rating is about to go up - this asset is among the best in the category - or down - this asset is among the worst in the category. In some ratings schemes, additional filters often provide lists which point out "the best purchase under $x", and so on. This observation does rely on smoothness in the sender and receiver distributions. Otherwise, it is possible to have adjacent pooling intervals with no separating zone in between; see Figure 3 for an illustration.

3.7. **Full Pooling.** Theorem 2 tells us very quickly when full pooling is optimal for the sender: if and only if S is first-order stochastically dominated by R. Our proposition is not stated using this wording since we must remember that S may not be a cdf.

Proposition 3. *Full pooling is optimal if and only if* $S(x) - S(0) \geq (1 - S(0))R(x)$ *for all* $x \in [\underline{a}, \bar{a}]$*. If* S *is a cdf with* $S(\underline{a}) = 0$*, then this condition is equivalent to* S *being first-order stochastically dominated by* R*.*

Proof. It is easy to see that, if $S(x) - S(0) \geq (1 - S(0))R(x)$ for all $x \in [\underline{a}, \bar{a}]$, then $\bar{a}$ is a successor of $\underline{a}$ and thus Theorem 2 implies full pooling is a solution to the school problem. In fact, let $a = \underline{a}$ and $a' = \bar{a}$ and check that (6) becomes $S(x) - S(0) \geq (1 - S(0))R(x)$. Conversely, if there exists x such that $S(x) - S(0) < (1 - S(0))R(x)$, then $\bar{a}$ is not a successor to $\underline{a}$ and, so by Theorems 1 and 2, can be improved upon. ■

3.8. **Full Separation.** We can also use Theorems 1 and 2 to determine when full revelation is optimal for the sender. Unlike the case of full pooling, first-order stochastic dominance is not the relevant criterion. Full revelation is optimal if and only if S conditional on any interval first-order stochastically dominates R conditional on any interval.

Proposition 4. *Full separation is optimal if and only if* $H = S \circ R^{-1}$ *is convex in* $[0, 1]$*. If* S *is a strictly increasing cdf, then this condition is equivalent to* $S(\cdot|(a, b))$ *first-order stochastically dominating* $R(\cdot|(a, b))$ *for every* $a, b \in [\underline{a}, \bar{a}]$*.*

Proof. The first statement is immediate given Theorem 1. As for the second, H is convex in $[0, 1]$ if and only if for every $w, x, z \in [0, 1]$ with $w < x \leqslant z$:

$$\frac{H(x) - H(w)}{x - w} \leqslant \frac{H(z) - H(w)}{z - w}$$

Letting $a = R^{-1}(w)$ and $b = R^{-1}(z)$ and $y = R^{-1}(x)$, this condition is equivalent to: for all $a, b, y \in [\underline{a}, \bar{a}]$ with $a < y \leqslant b$,

$$\frac{S(y) - S(a)}{R(y) - R(a)} \leqslant \frac{S(b) - S(a)}{R(b) - R(a)}$$

If S is a strictly increasing cdf, then this condition is equivalent to $S(\cdot|(a, b))$ first-order stochastically dominating $R(\cdot|(a, b))$ for every $a, b \in [\underline{a}, \bar{a}]$. ■

4. Application: Educational Grades

The design of quality categories when a sender and a receiver have different priors is a problem that arises in a number of applications. Most obviously, think of a seller who is better informed about the goods being sold than buyers. Our results say that, unless the seller knows that the distribution of quality of her products is first-order stochastically dominated by the prior of the consumer, she would like to categorize her goods.

Or think of a retailing intermediary who charges as a fee a proportion of the value of a sale. Suppose that this intermediary receives different fees for different products. These differences in fees translate into a change in the distribution the intermediary cares about. She might then design categories so as to attain a higher price for the goods for which she gets a higher fee. In the same vein, we can think of a public authority designing affirmative action through a grading system. The authority can design the grading of a public test so as to maximize the outcome of the underprivileged population of interest.

Apart from these applications in which the sender and receiver naturally hold different priors, the solution to an additional moral hazard problem can effectively map into a distortion of the sender's distribution. The design of educational grades is one such problem, and it is the application we have chosen to highlight.

4.1. **School and Students.** A private university — our sender — is designing a grading system for its students with a sharp eye on its tuition revenues, which it seeks to maximize. This is part of a larger problem in which it might choose an admissions cutoff. We ignore that problem and suppose that it has already chosen its student body.

A mass of students of unit measure enter the university, with abilities distributed according to a cdf R on $[\underline{a}, \bar{a}]$, where $\underline{a}$ is the admissions cutoff. Abilities have value to potential employers, and also affect learning costs as in a standard Spence model. Upon entering, a student learns her ability and chooses a learning level. A student with ability a who learns ℓ incurs a cost $c(a)\ell$, where $c'(a) < 0$. But prior to all that, students must choose whether to enter school. This decision is often made before ability is fully learnt, though it may be known to a greater or lesser degree. Specifically, at this anterior stage the student has a belief about her own ability, represented as a distribution over $[\underline{a}, \bar{a}]$. That belief may involve full knowledge (a degenerate distribution), or just the population belief R, or it may be some other distribution.[6] In short, we have a set of *belief types* represented by some collection of prior beliefs on ability. Among these, we assume that there is some belief F_0 such as that all other belief types dominate F_0 according to first-order stochastic dominance. We assume that both R and F_0 are continuous cdfs.

[6]Before joining the school, students don't know where they will place in the distribution of outcomes, but once they join the school and their type is realized, everything is deterministic. We make this modeling decision in order to avoid having a contest with stochastic outcomes.

4.2. **Job Market.** A market (our receiver) employs all students who leave the school. The market does not directly observe student abilities, but infers these from a school-provided signal: the student's transcript of grades or her *learning*, which we proxy by the observable scalar ℓ. The market values one unit of ability by one unit of money (a harmless normalization), and each unit of learning by λ, and so pays out

$$\mathbb{E}(a|\ell) + \lambda\ell \tag{7}$$

to a student whom they know has a learning of ℓ, where the conditional expectation will be determined by the distribution R as well as equilibrium learning strategies.

4.3. **Categorization.** The school chooses a learning technology: a compact set of certified learning levels L. A student of ability a chooses $\ell \in L$ by exerting effort at cost $c(a)\ell$. Leaning nothing is always an option, so $0 \in L$. The student *could* choose $\ell \notin L$ but the market doesn't observe this; rather, it "sees" only the highest certified level consistent with ℓ. That is, the market sees $\ell' = \max\{\ell'' \in L | \ell'' \leqslant \ell\}$. With this restriction in hand the student must effectively choose from the certified learning levels on offer.

For the rest of the paper, we assume that market rewards for learning *alone* are not enough to induce learning by even the highest ability (lowest cost) student:

[C] $c(\bar{a}) > \lambda$.

4.4. **Incentive-Compatible Learning.** By using the language of direct mechanisms we may as well imagine that the school chooses some incentive compatible learning function $\ell : [\underline{a}, \bar{a}] \to \mathbb{R}_+$ assigning a learning level to each ability type. To save on unnecessary notation, we presume that the school always chooses L so that every non-zero learning level is occupied by *some* ability type. The only learning level for which this presumption is *with* loss of generality is $\ell = 0$, and we always retain that option for the student, though the school may not want to implement it for any of its admitted types. We can now characterize the set of incentive compatible learning functions. For a given learning function ℓ, the value to an obeying student of ability a is

$$\mathbb{E}_R(a'|\ell(a)) + \lambda\ell(a) - c(a)\ell(a)$$

where $\mathbb{E}_R(a'|\ell(a))$ is the expectation of ability a' given that $\ell(a)$ is observed, and given that all students follow ℓ. A standard single-crossing argument yields:

Lemma 1. *If $\ell(a)$ and $\ell(a')$ are optimal for a and a', with $a > a'$, then $\ell(a) \geq \ell(a')$.*

Remember that $\mathcal{A}_R$ is the collection of all categorizations satisfying (1). For every ℓ, there is a categorization $A_\ell \in \mathcal{A}_R$ which is a quality estimate function associated to ℓ. On a strictly increasing patch of ℓ, abilities are fully revealed: $A_\ell(a) = a$. On a constant patch, the ability estimate equals the appropriate conditional mean.

To ensure that every student willingly abides by ℓ — that is, she chooses the learning level $\ell(a)$ when her ability is a — we impose the incentive compatibility condition: for every $a, a' \in [\underline{a}, \bar{a}]$,

$$A_\ell(a) + \lambda\ell(a) - c(a)\ell(a) \geq \max\{A_\ell(a') + \lambda\ell(a') - c(a)\ell(a'), 0\} \tag{8}$$

The nonnegativity constraint is explicitly specified to handle the case $\ell(\underline{a}) > 0$. Then the zero-learning choice is off-path, and suitable beliefs will be needed to guarantee incentive-compatibility. We presume that in such cases, the observation of 0 is associated with the belief that the student has the lowest ability $\underline{a}$. This implements the best equilibrium from the perspective of a tuition-maximizing school.[7]

We now describe incentive-compatible learning technologies in more detail. For any nondecreasing function f, let $f^{\uparrow}(x)$ stand for the left-hand limit of f at x.

Proposition 5. *A learning function $\ell(a)$ is incentive-compatible if and only if*

(i) *It is nondecreasing with $A_\ell(\underline{a}) + \lambda\ell(\underline{a}) - c(\underline{a})\ell(\underline{a}) \geqslant 0$.*

(ii) *It is differentiable on any increasing patch, with*

$$\ell'(a) = \frac{1}{c(a) - \lambda}; \tag{9}$$

(iii) *For any threshold t dividing two adjacent patches,*

$$\ell(t) = \ell^{\uparrow}(t) + \frac{A_\ell(t) - A_\ell^{\uparrow}(t)}{c(t) - \lambda}. \tag{10}$$

so that — like A_ℓ — ℓ must jump up across patches.

Proposition 6. *For every $A \in \mathcal{A}_R$, there is an incentive-compatible ℓ, unique up to the choice of $\ell(\underline{a}) \in \left[0, \frac{A(\underline{a})}{c(\underline{a})-\lambda}\right]$, satisfying* (i) *and* (ii) *of Proposition 5, such that $A_\ell = A$.*

We conclude that the choice of an incentive compatible learning technology is equivalent to choosing its pooling and separating patches, up to an initial constant determined by the learning $\ell(\underline{a})$ assigned to a zero-ability individual. For more on this initial assignment, see Section 4.6. In Figure 4, we draw an example of pooling and separating patches and the A_ℓ and ℓ functions associated with them. We use this connection between A and ℓ to redefine the school problem in the next section and solve for the optimal mechanism.

4.5. **Tuition.** The school sets one tuition level. The less optimistic a student is about his ability, the less she is willing to pay for schooling. Belief type F_0 has the lowest willingness to pay, so to extract maximal revenue, the school maximizes the expected payoff of this type, not counting tuition fees. As already mentioned, this can be viewed

[7]Obviously, any other belief will impose tighter constraints on the school design problem.

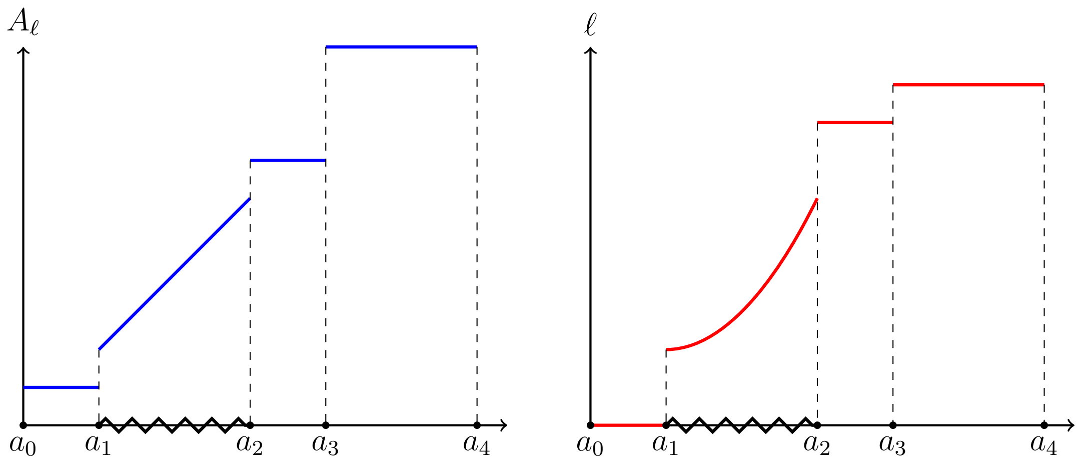


FIGURE 4. Example of a market ability estimate function $A_\ell \in \mathcal{A}_R$ and incentive compatible learning function ℓ associated to the pooling patches $[a_0, a_1)$, $[a_2, a_3)$ and $[a_3, a_4]$ and separating patch $[a_1, a_2)$. The separating patch is indicated by the jagged line.

as a second stage of the school's decision process, wherein it first chooses which types to serve and then the tuition level and testing technology, given the types it has chosen to serve. The profit-maximizing tuition price must then be so that this lowest belief type is indifferent between joining the school and its next best option, normalized to zero.

4.6. **School Payoffs.** Under some incentive compatible learning function ℓ, the net payoff to a student of ability a is given by

$$A_\ell(a) + [\lambda - c(a)]\,\ell(a).$$

The profit to the school equals the tuition level it charges, which we argued is equal to the value it yields to the lowest belief type in the economy. Hence, the school's objective is fully aligned with the objective of the student with the lowest belief type. In order to maximize profits, the school must then choose an incentive compatible learning function that maximizes the ex-ante payoff of the student with beliefs F_0, which is given by

$$\int_{\underline{a}}^{\bar{a}} [A_\ell(a) + \{\lambda - \sigma c(a)\}\ell(a)]dF_0(a), \tag{11}$$

where $\sigma \in [0, 1]$ is the extent to which the student (or parent) internalizes effort costs at the ex-ante stage.[8] The extent of this internalization will determine not just the level of

[8]An interesting question is the degree to which this tuition is affected by the cost of effort while in school. One possibility is that tuition is paid by a parent who values the child's wage prospects but does not internalize the cost of effort at all. The other extreme is that the student pays the tuition and knows what she is in for. We remain agnostic and suppose that a fraction $\sigma \in [0, 1]$ of the expected learning cost is internalized at the time of tuition payment.

the tuition (which is a relatively minor consideration, at least for the analysis), but also the school's "attitude" towards the intrinsic value of learning; more on this below.

4.7. **Solution to the School Problem.** We now link the school problem to our more abstract setting, thereby permitting a full solution of it.

Proposition 7. *Assume Condition C. For every $A \in \mathcal{A}_R$ and associated ℓ with $\ell(\underline{a}) = \frac{A(\underline{a})}{c(\underline{a})-\lambda}$ as described in Proposition 6,*

$$\int_{\underline{a}}^{\bar{a}} [A(a) + \{\lambda - \sigma c(a)\}\ell(a)]\, dF_0(a) = \int_{\underline{a}}^{\bar{a}} A(a) dS(a) \tag{12}$$

where

$$S(a) = F_0(a) + \int_a^{\bar{a}} \frac{\sigma c(x) - \lambda}{c(a) - \lambda} dF_0(x) \tag{13}$$

is a continuous function of bounded variation with $S(\underline{a})$ finite and $S(\bar{a}) = 1$, so that the Stieltjes integral in (2) is well-defined.

Proposition 7 achieves a significant simplification. It asserts that for all learning functions for which the incentive constraint (8) at $\underline{a}$ binds, the school payoff is simply equal to $\int_{\underline{a}}^{\bar{a}} A(a)dS(a)$, where S is defined by (13). Therefore all we need to do to solve the school problem in this case is to apply Theorem 2 for the categorization problem, with receiver distribution R and sender distribution equal to the constructed function S.

What about the complications of moral hazard associated with learning, and its associated incentive constraints? And what about the intrinsic payoffs from learning? All of that gets fully subsumed in the function S, so in the end S can be viewed as a "distortion" of students' ex-ante beliefs F_0, as equation (13) makes clear. So this example reveals, at one stroke, two reasons for the "prior" S to be different from the distribution R, neither of which have anything to do with the absence of common priors between sender and receiver. One is that the sender may effectively be delegated to work on behalf of someone with a distinct prior, as she is pushed to do here in the interests of profit-maximization. The other stems from additional distortions caused by incentive constraints or by payoffs from related sources. These latter considerations can even cause S to depart from a cdf.

The school problem can therefore be seen as an instance of our abstract categorization problem, at least when the school wants to generate enough learning so that the incentive constraint (8) binds. As discussed below, that will happen when effort costs are only internalized to a small extent (see (15). Otherwise, if the school internalizes effort costs enough, it cuts back on learning in its effort to maximize tuition revenues. The resulting solution must then go beyond Proposition 7. It turns out that a simple auxiliary step permits a complete solution:

Theorem 3. *Assume Condition C. Define* $\Theta = \max\left\{0, \int_{\underline{a}}^{\bar{a}} \frac{\lambda - \sigma c(x)}{c(\underline{a}) - \lambda} dF_0(x)\right\}$.

[I] *If* $\Theta = 0$*,* A^* *is the solution to the optimal categorization problem, with* R *as the receiver's distribution and* S*, as defined in* (13)*, as the sender's distribution. The optimal learning function is the unique* ℓ *associated with* A^* *with* $\ell(\underline{a}) = \frac{A^*(\underline{a})}{c(\underline{a})-\lambda}$.

[II] *If* $\Theta > 0$*, then the solution can be found by a two step procedure:*

Step 1*. Solve the categorization problem in the set* $[\alpha, \bar{a}]$ *for every* $\alpha \in [\underline{a}, \bar{a}]$*, with* R *as the receiver's distribution conditional on* $[\alpha, \bar{a}]$*, and* S*, as defined in* (13) *without any conditioning, as the sender's distribution. Let* $A_\alpha \in \mathcal{A}_R$ *be the categorization that pools every type in* $[\underline{a}, \alpha)$ *and follows the optimal categorization on* $[\alpha, \bar{a}]$*.*

Step 2*. Find* $\alpha \in [\underline{a}, \bar{a}]$ *that maximizes*

$$\int_{\underline{a}}^{\bar{a}} A_\alpha(x) dS(x) + \Theta A_\alpha(\underline{a}) \tag{14}$$

Let A^* *be the categorization associated with the* α *that solves this problem. The optimal learning function is the unique* ℓ *associated with* A^* *such that* $\ell(\underline{a}) = 0$*.*

To understand the theorem, recall the parameter σ, which can be interpreted as the extent to which the student (or parent) — and therefore, by extension, the profit-making school — internalizes the cost of learning. If

$$\lambda \geqslant \sigma \int_{\underline{a}}^{\bar{a}} c(a) dF_0(a), \tag{15}$$

then on average under F_0, parents value the payoff λ from one more unit of learning, compared to the expected cost of that learning, multiplied by the internalization parameter σ. In this case, the school would like to shift its entire learning function upward, so that all types under the support of F_0 can benefit from the extra income — and the school can charge higher tuition. Of course, there's a limit to that: the student's incentive constraint in (8) will bind, determining the value of $\ell(\underline{a})$ as given in Part I of the Theorem. Now Proposition 7 kicks in, permitting the school problem to be directly converted into the categorization problem via the function S, as already discussed.

On the other hand, when (15) fails, then the learning component of future income will not allow the school to earn additional tuition revenue, as the internalized effort costs of acquiring that learning are too high. In this case, it will be optimal to set $\ell(\underline{a}) = 0$, and the incentive constraint (8) fails to bind. This failure has a technical implication of conceptual import: the mutual coherence of the two mappings ℓ and A cannot be invoked to generate the equality of integrals in (13); see Lemma 3 in Section 5 for a more precise understanding of this point. In turn, this prevents an immediate conversion of the school problem into the optimal categorization problem. However, as Part II of the Theorem shows, it does not eliminate that possibility either. We need to remove the one non-binding incentive constraint, and we do so by probing the optimal categorization problem for *each* choice α of the initial pooling category, as described by the intervals $[\underline{a}, \alpha]$. With that done, we choose α as described in Step 2 to solve the school problem.

4.8. **Pooling.** The following results apply Theorems 2 and 3 and Proposition 3. The first tells us that a school will want to pool all ability types whenever the market places a low relative value on learning, as in Lizzeri (1999), where learning has no value. The second states that if students fully anticipate their cost of learning ex-ante, then the school to fully pool all students. Finally, if the lowest belief student is certain ex-ante that she is the lowest ability type, then pooling all types is also profit-maximizing for the school.

Proposition 8. *A sufficient condition for full pooling to be optimal is:*

$$\int_{a}^{\bar{a}} [\sigma c(x) - \lambda] dF_0(x) \geqslant 0 \quad \textit{for every } a \in [\underline{a}, \bar{a}] \tag{16}$$

(i) (16) *is satisfied when* $\lambda = 0$*. Moreover, if there is* $\lambda > 0$ *such that it is satisfied, then it is also satisfied for any* $\lambda' < \lambda$*.*

(ii) (16) *is satisfied when* $\sigma = 1$*. Moreover, if there is* $\sigma > 0$ *such that it is satisfied, then it is also satisfied for any* $\sigma' > \sigma$*.*

(iii) *If there is* F_0 *such that* (16) *is satisfied, then it is also satisfied for any* F_0' *such that* F_0 *first order stochastically improves over* F_0'*.*

The takeaway from these results is that the solution to the school problem only involves any separation if learning is valued enough by the market, if learning is not too costly ex-ante, and if the lowest belief-type agent is optimistic enough about their ability type.

4.9. **A Solution with Some Pooling and Some Separation.** We apply our results to a particular problem. Let R be uniform on $[0, 1]$, and $F_0 = a^\gamma$ for $\gamma \in [0, 1]$. If $\gamma = 1$, then $F_0 = R$: students have no ex-ante information about their own abilities. For lower γ, F_0 is first-order stochastically dominated by R, so that the lowest types are pessimistic relative to the population average. The lower is γ, the further the belief of the most pessimistic type from that of the average agent. Set $c(a) = 1/a$ and $\sigma = 0$. This last restriction means that the cost of effort is not internalized by the students (or their parents) when choosing to join the school. To comply with Condition C, we set $\lambda < 1$.

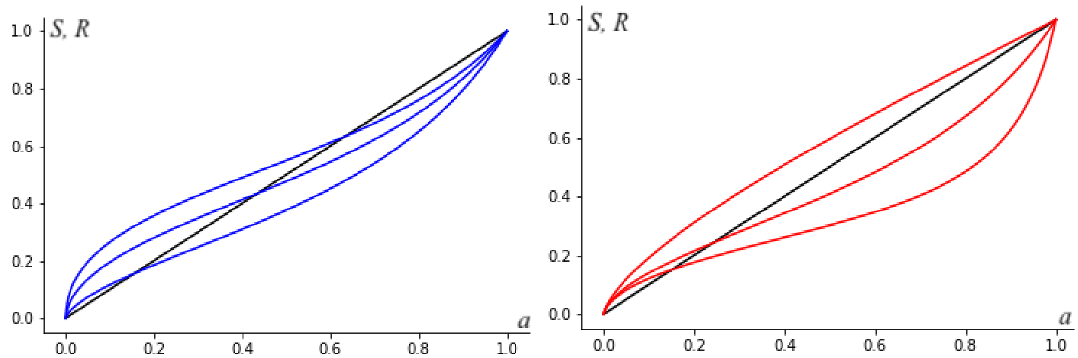


FIGURE 5. In both panels, the uniform distribution R is represented in black. In the left panel, S is represented in blue for three different values of γ: S increases in the first-order stochastic dominance ordering as γ increases. In the right panel, S is represented in red for three different values of λ: again, S increases in the first-order stochastic dominance ordering as λ increases.

With these functional forms, use equation (13) to map the school's "distorted prior" S:

$$S(a) = \frac{a^\gamma - \lambda a}{1 - \lambda a} \tag{17}$$

In this special case, S is a cdf (see Figure 5), and Case I of Theorem 3 applies. So, following the procedure described in the previous section, we can fully pin down the signaling structure that solves the school problem, as stated in the proposition below.

Proposition 9. *There is $\tilde{a} \in (0, 1]$ such that the solution to the school problem is to pool agents with $a \in [0, \tilde{a})$, and fully reveal the ability of all agents with $a \geq \tilde{a}$.*

Proof. We show that either 1 is the largest successor of 0 or there is an $\tilde{a} \in (0, 1)$ such that $\tilde{a}$ is the largest successor of 0 and no $a \in [\tilde{a}, 1]$ has a successor. Then, Proposition 1 implies the result.

Towards finding the largest successor of 0, let $h(a) = \frac{S(a)}{R(a)}$ and take a derivative of h with respect to a to find

$$h(a) = \frac{a^{\gamma-1} - \lambda}{1 - \lambda a}$$

$$h'(a) = \frac{(\gamma - 1)a^{\gamma-2}}{1 - \lambda a} + \frac{\lambda h(a)}{1 - \lambda a} \tag{18}$$

For a small, $h'(a) < 0$, since $\gamma < 1$. Now suppose there is $\hat{a} \in (0, 1]$ such that $h'(\hat{a}) \geq 0$. Then $h'(a) \geq 0$ for all $a \in [\hat{a}, 1]$, because the first term in the numerator of (18) is negative, but diminishes in magnitude as a increases, while the second term is positive and, if $h' \geq 0$, gets more positive as a increases.

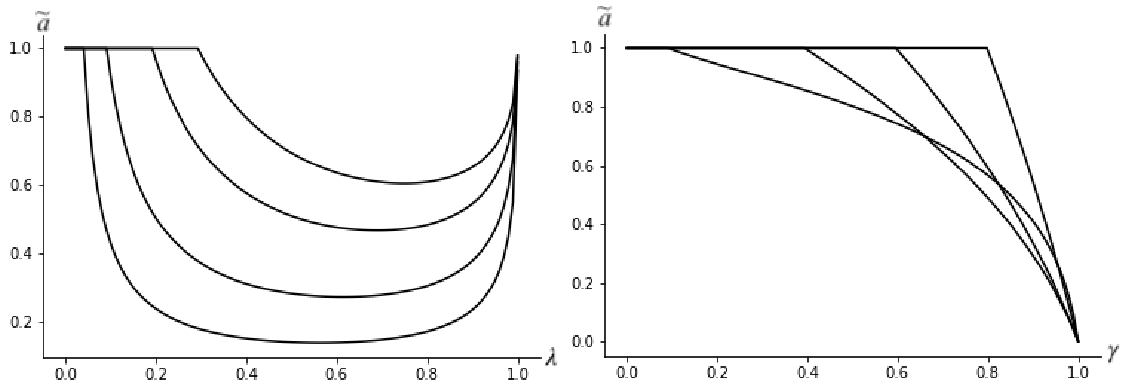


FIGURE 6. The left panel plots $\tilde{a}$ as a function of λ for four different values of γ — lower lines correspond to higher γ. The right panel plots $\tilde{a}$ as a function of γ for four different values of λ — lower lines correspond to higher λ.

So there exists $\tilde{a} \in (0, 1]$ such that h is decreasing in a for $a \leqslant \tilde{a}$ and increasing in a for $a > \tilde{a}$. It then follows, by the definition of successor, that $\tilde{a}$ is the largest successor of 0.

Now we can check that if $h'(a) \geqslant 0$, then $S''(a) \geqslant 0$, and so S is convex in $[\tilde{a}, 1]$. But this implies that no $a \in [\tilde{a}, 1]$ has a successor. ■

Figure 6 depicts $\tilde{a}$ as a function of λ and γ. When the lowest-belief type is more optimistic, i.e. γ is higher, $\tilde{a}$ is lower and there is more separation. The connection with the market value for learning is not monotonic. Initially, higher value for learning induces more separation, but for high values of λ, increasing λ leads to more pooling.

5. Omitted Proofs and Details

Proof of Observation 1. Because $A \in \mathcal{A}_R$ and R is strictly increasing, we have:

$$\begin{aligned}
\int_{\underline{a}}^{\bar{a}} A(a)dS(a) &= \sum_{s\in S}\int_s^{s'} adS(a) + \sum_{p\in P}\mathbb{E}_R\left(a|a\in[p,p')\right)\left[S(p')-S(p)\right] \\
&= \sum_{s\in S}\int_s^{s'} adS(a) + \sum_{p\in P}\int_p^{p'} adR(a)\left[\frac{S(p')-S(p)}{R(p')-R(p)}\right] \\
&= \sum_{s\in S}\int_s^{s'} ad\Psi(a,A) + \sum_{p\in P}\int_p^{p'} ad\Psi(a,A) = \int_{\underline{a}}^{\bar{a}} ad\Psi(a,A),
\end{aligned}$$

where in the penultimate step $d\Psi$ is well defined as Ψ has bounded variation, and the last equality follows from the continuity of Ψ in a. ■

Proof of Observation 4. Parts (i) and (ii) follows immediately from the continuity of R and S. To show part (iii), we must prove that

$$\frac{S(x)-S(a)}{R(x)-R(a)} \geq \frac{S(b')-S(a)}{R(b')-R(a)} \tag{19}$$

for every $x \in (a, b']$. We establish (19) in three steps.

Step 1: $\dfrac{S(b)-S(a)}{R(b)-R(a)} \geq \dfrac{S(b')-S(b)}{R(b')-R(b)}$.

Proof. Because $a' \succ a$ and $b \in [a, a']$, we have:

$$\frac{S(b)-S(a)}{R(b)-R(a)} \geq \frac{S(a')-S(a)}{R(a')-R(a)}. \tag{20}$$

At the same time, by elementary accounting,

$$\frac{S(a')-S(a)}{R(a')-R(a)} = \frac{R(b)-R(a)}{R(a')-R(a)}\left[\frac{S(b)-S(a)}{R(b)-R(a)}\right] + \frac{R(a')-R(b)}{R(a')-R(a)}\left[\frac{S(a')-S(b)}{R(a')-R(b)}\right] \tag{21}$$

Combine (20) and (21) and rearrange to get

$$\frac{S(b)-S(a)}{R(b)-R(a)} \geq \frac{S(a')-S(b)}{R(a')-R(b)}. \tag{22}$$

Because $b' \succ b$ and $a' \in [b, b']$,

$$\frac{S(a')-S(b)}{R(a')-R(b)} \geq \frac{S(b')-S(b)}{R(b')-R(b)} \tag{23}$$

Together, (22) and (23) imply Step 1.

Step 2: Inequality (19) holds for all $x \in (a, a']$.

Proof. Because $b' \succ b$ and $a' \in [b, b']$,

$$\frac{S(a')-S(b)}{R(a')-R(b)} \geq \frac{S(b')-S(b)}{R(b')-R(b)} \tag{24}$$

Invoking the accounting identity (21), and combining it with (24), we get:

$$\begin{aligned}
\frac{S(a')-S(a)}{R(a')-R(a)} &= \frac{R(b)-R(a)}{R(a')-R(a)}\left[\frac{S(b)-S(a)}{R(b)-R(a)}\right]+\frac{R(a')-R(b)}{R(a')-R(a)}\left[\frac{S(a')-S(b)}{R(a')-R(b)}\right]\\
&\geq \frac{R(b)-R(a)}{R(a')-R(a)}\left[\frac{S(b)-S(a)}{R(b)-R(a)}\right]+\frac{R(a')-R(b)}{R(a')-R(a)}\left[\frac{S(b')-S(b)}{R(b')-R(b)}\right]\\
&= \frac{R(b)-R(a)}{R(a')-R(a)}\left[\frac{S(b)-S(a)}{R(b)-R(a)}\right]+\left[1-\frac{R(b)-R(a)}{R(a')-R(a)}\right]\left[\frac{S(b')-S(b)}{R(b')-R(b)}\right]\\
&\geq \frac{R(b)-R(a)}{R(b')-R(a)}\left[\frac{S(b)-S(a)}{R(b)-R(a)}\right]+\left[1-\frac{R(b)-R(a)}{R(b')-R(a)}\right]\left[\frac{S(b')-S(b)}{R(b')-R(b)}\right]\\
&= \frac{S(b')-S(a)}{R(b')-R(a)}, \qquad (25)
\end{aligned}$$

where the first inequality uses (24), and the second inequality uses Step 1 along with the fact that $\frac{R(b)-R(a)}{R(b')-R(a)} \leqslant \frac{R(b)-R(a)}{R(a')-R(a)}$. At the same time, given $a' \succ a$,

$$\frac{S(x)-S(a)}{R(x)-R(a)} \geq \frac{S(a')-S(a)}{R(a')-R(a)} \quad \text{for every } x \in (a, a']. \qquad (26)$$

Together, (25) and (26) imply Step 2.

Step 3: Inequality (19) holds for all $x \in (a', b']$.

Proof. Because $b' \succ b$ and $x > a' > b$,

$$\frac{S(x)-S(b)}{R(x)-R(b)} \geq \frac{S(b')-S(b)}{R(b')-R(b)} \qquad (27)$$

Using (27) in an elementary accounting identity, we therefore see that

$$\begin{aligned}
\frac{S(x)-S(a)}{R(x)-R(a)} &= \frac{R(b)-R(a)}{R(x)-R(a)}\left[\frac{S(b)-S(a)}{R(b)-R(a)}\right]+\frac{R(x)-R(b)}{R(x)-R(a)}\left[\frac{S(x)-S(b)}{R(x)-R(b)}\right]\\
&\geq \frac{R(b)-R(a)}{R(x)-R(a)}\left[\frac{S(b)-S(a)}{R(b)-R(a)}\right]+\frac{R(x)-R(b)}{R(x)-R(a)}\left[\frac{S(b')-S(b)}{R(b')-R(b)}\right]\\
&\geq \frac{R(b)-R(a)}{R(b')-R(a)}\left[\frac{S(b)-S(a)}{R(b)-R(a)}\right]+\frac{R(b')-R(b)}{R(b')-R(a)}\left[\frac{S(b')-S(b)}{R(b')-R(b)}\right]\\
&= \frac{S(b')-S(a)}{R(b')-R(a)}, \qquad (28)
\end{aligned}$$

where the first equality is just accounting, the first inequality invokes (27), and the second inequality uses both Step 1 and $\frac{R(b)-R(a)}{R(x)-R(a)} \geq \frac{R(b)-R(a)}{R(b')-R(a)}$. (28) establishes Step 3.

Steps 2 and 3 together imply (19) for all $x \in (a, b']$, so $b' \succ a$. ∎

Lemma 2. *For any maximal interval with edges p and p', define the corresponding percentiles $q = R(p)$ and $q' = R(p')$. Then for all $z \in [q, q']$, recalling that $H = S \circ R^{-1}$,*

$$H(z) \geqslant \Phi(z, A^*) = H(q) + (z - q)\left[\frac{H(q') - H(q)}{q' - q}\right]. \tag{29}$$

Proof. By (3), we know that for every $x \in [p, p']$,

$$\Psi(x, A^*) = S(p) + [R(x) - R(p)]\left[\frac{S(p') - S(p)}{R(p') - R(p)}\right]. \tag{30}$$

Because $p' \succ p$, we have $\frac{S(x)-S(p)}{R(x)-R(p)} \geqslant \frac{S(p')-S(p)}{R(p')-R(p)}$, or

$$S(x) \geqslant S(p) + [R(x) - R(p)]\left[\frac{S(p') - S(p)}{R(p') - R(p)}\right]. \tag{31}$$

Convert to percentiles using $z = R(x)$ for any generic element $x \in [p, p']$, and combine (30) and (31) to obtain (29), as desired. ∎

Proof of Theorem 2. Let A^* be as described in the statement of the theorem. Observation 4 assures us that every maximal interval is well-defined, and therefore so is A^*. By Theorem 1, it will suffice to prove that $\Phi(z, A^*) = \breve{H}(z)$ for all $z \in [0, 1]$. This is obviously true for $z = 0$ and $z = 1$. Suppose, on the contrary, that $\Phi(\hat{z}, A^*) \neq \breve{H}(\hat{z})$ for some $z \in (0, 1)$. Then, by Observation 2,

$$\Phi(\hat{z}, A^*) > \breve{H}(\hat{z}). \tag{32}$$

We observe that $H(\hat{z}) > \breve{H}(\hat{z})$. If not, then by (32), we have $\Phi(\hat{z}, A^*) > H(\hat{z})$, which contradicts Lemma 2. So $H(\hat{z}) > \breve{H}(\hat{z})$, and in particular, $\breve{H}(\hat{z})$ must lie on a line segment that connects two points on the graph of H and lies below H in between. Specifically, there is w and w', both in $[0, 1]$ with $w < \hat{z} < w'$, $H(w) = \breve{H}(w)$, $\breve{H}(w') = w'$, and for every $z \in (w, w')$, including $\hat{z}$,

$$\frac{H(z) - H(w)}{z - w} \geqslant \frac{\breve{H}(z) - H(w)}{z - w} = \frac{H(w') - H(w)}{w' - w}. \tag{33}$$

Define $t \equiv R^{-1}(w)$, $t' \equiv R^{-1}(w')$, $\hat{x} \equiv R^{-1}(\hat{z})$, and $x \equiv R^{-1}(z)$ for generic $z \in (w, w')$. Then $\hat{x} \in (t, t')$, and recalling that $H = S \circ R^{-1}$, (33) implies that

$$\frac{S(x) - S(t)}{R(x) - R(t)} \geqslant \frac{S(t') - S(t)}{R(t') - R(t)}$$

for all $x \in (t, t')$, which proves that $t' \succ t$, so that $[t, t')$ must be a weak subset of some maximal interval. Consequently, by (3) and the construction of A^*,

$$\Psi(x, A^*) = S(p) + [R(x) - R(p)]\left[\frac{S(p') - S(p)}{R(p') - R(p)}\right] \tag{34}$$

for some maximal interval $[p, p')$. Because $\hat{x} \in [t, t')$ and $\hat{x} \in [p, p')$, these intervals are not disjoint, but because $[p, p')$ is maximal, Observation 4 (iii) implies that

$$[t, t') \subseteq [p, p'). \tag{35}$$

In particular, because $p' \succ p$, this means that

$$\frac{S(t) - S(p)}{R(t) - R(p)} \geqslant \frac{S(p') - S(p)}{R(p') - R(p)} \leqslant \frac{S(t') - S(p)}{R(t') - R(p)}. \tag{36}$$

Now we rewrite these relationships in percentile format. To this end, let $q = R(p)$ and $q' = R(p')$, then (34) becomes

$$\Phi(z, A^*) = H(q) + (z - q) \left[\frac{H(q') - H(q)}{q' - q} \right] = H(q) + (z - q)\alpha, \tag{37}$$

for all $z \in [q, q')$, where we define $\alpha \equiv \frac{H(q') - H(q)}{q' - q}$. Inequality (36) then becomes $\frac{H(w) - H(q)}{w - q} \geqslant \alpha \leqslant \frac{H(w') - H(q)}{w' - q}$, or equivalently

$$H(w) \geqslant H(q) + (w - q)\alpha \text{ and } H(w') \geqslant H(q) + (w' - q)\alpha, \tag{38}$$

Using (33), we must conclude that

$$\breve{H}(\hat{z}) \geqslant H(q) + (\hat{z} - q)\alpha = \Phi(z, A^*), \tag{39}$$

but (39) contradicts (32), and so completes the proof.

Proof of Proposition 2. Suppose by contradiction that $[a, b)$ and $[b, c)$ with $a < b < c$ are two adjacent *maximal* pooling intervals, with no separating interval in between.

Claim 1. $\dfrac{S(x) - S(b)}{R(x) - R(b)} \geqslant \dfrac{S(c) - S(b)}{R(c) - R(b)} > \dfrac{S(b) - S(a)}{R(b) - R(a)}$ for every $x \in (b, c)$.

Proof. The weak inequality is true because $[b, c)$ is pooling and so $c \succ b$. We show the strict inequality. If it is false, then

$$\frac{S(b) - S(a)}{R(b) - R(a)} \geqslant \frac{S(c) - S(b)}{R(c) - R(b)} \geqslant \frac{S(c) - S(a)}{R(c) - R(a)}, \tag{40}$$

where the second inequality is a consequence of the first. Combining (40) with the fact that $b \succ a$, we see that for any $x \in (a, b]$,

$$\frac{S(x) - S(a)}{R(x) - R(a)} \geqslant \frac{S(c) - S(a)}{R(c) - R(a)}. \tag{41}$$

Moreover, for any $x \in (b, c)$,

$$\begin{aligned}
(42) \qquad \frac{S(x)-S(a)}{R(x)-R(a)} &= \frac{R(x)-R(b)}{R(x)-R(a)}\frac{S(x)-S(b)}{R(x)-R(b)} + \frac{R(b)-R(a)}{R(x)-R(a)}\frac{S(b)-S(a)}{R(b)-R(a)} \\
&\geqslant \frac{R(x)-R(b)}{R(x)-R(a)}\frac{S(c)-S(b)}{R(c)-R(b)} + \frac{R(b)-R(a)}{R(x)-R(a)}\frac{S(b)-S(a)}{R(b)-R(a)} \\
&\geqslant \frac{R(c)-R(b)}{R(c)-R(a)}\frac{S(c)-S(b)}{R(c)-R(b)} + \frac{R(b)-R(a)}{R(c)-R(a)}\frac{S(b)-S(a)}{R(b)-R(a)} = \frac{S(c)-S(a)}{R(c)-R(a)},
\end{aligned}$$

where the first inequality follows from $c \succ b$ and the second one from (40). But (41) and (42) together imply that $c \succ a$ is a maximal pair, which contradicts the fact $[a, b)$ is maximal. This contradiction establishes the Claim. ■

Claim 2. $\dfrac{S(b)-S(x)}{R(b)-R(x)} \leqslant \dfrac{S(b)-S(a)}{R(b)-R(a)}$ for every $x \in (a, b)$.

Proof. Because $[a, b)$ is maximal,

$$\frac{S(x)-S(a)}{R(x)-R(a)} \geqslant \frac{S(b)-S(a)}{R(b)-R(a)} \tag{43}$$

for every $x \in (a, b)$. Moreover,

$$\frac{S(b)-S(a)}{R(b)-R(a)} = \frac{R(x)-R(a)}{R(b)-R(a)}\frac{S(x)-S(a)}{R(x)-R(a)} + \frac{R(b)-R(x)}{R(b)-R(a)}\frac{S(b)-S(x)}{R(b)-R(x)} \tag{44}$$

(43) and (44) together imply the claim. ■

Let r and s be the densities of R and S respectively. Because Claim 1 establishes a uniform lower bound on $\frac{S(x)-S(b)}{R(x)-R(b)}$, we must conclude that

$$\lim_{x \to b} \frac{S(x)-S(b)}{R(x)-R(b)} = \frac{s(b)}{r(b)} > \frac{S(b)-S(a)}{R(b)-R(a)}. \tag{45}$$

But Claim 2 implies that

$$\lim_{x \to b} \frac{S(b)-S(x)}{R(b)-R(x)} = \frac{s(b)}{r(b)} \leqslant \frac{S(b)-S(a)}{R(b)-R(a)},$$

which contradicts (45). ■

Proof of Proposition 5. Let $\ell(a)$ be an incentive-compatible learning function. By Lemma 1, ℓ is nondecreasing. Moreover, $A_\ell(\underline{a}) + \lambda\ell(\underline{a}) - c(\underline{a})\ell(\underline{a}) \geqslant A \geqslant 0$, where A is the (possible off-path) ability estimate when 0 is observed.

Next, take a and a' in the same increasing patch. From (1) and (8), we have:

$$\frac{1}{c(a') - \lambda} \geq \frac{\ell(a) - \ell(a')}{a - a'} \geq \frac{1}{c(a) - \lambda}.$$

Take $a' \to a$ to obtain (9) on any increasing patch.

Now take $a \in [a_{k-1}, a_k)$ for some $k > 1$. Use (8) to see that

$$\ell(a) + \frac{A_\ell(a_k) - A_\ell(a)}{c(a) - \lambda} \leq \ell(a_k) \leq \ell(a) + \frac{A_\ell(a_k) - A_\ell(a)}{c(a_k) - \lambda},$$

and take $a \to a_k$ to get (10). Now note that $\ell(a)$ is constant on non-increasing patches to conclude that (i) and (ii) are also sufficient for an incentive-compatible learning function $\ell(a)$. ■

Proof of Proposition 6. Let $A \in \mathcal{A}_R$. For any $\ell(\underline{a}) \in \left[0, \frac{A(\underline{a})}{c(\underline{a})-\lambda}\right]$, define a function ℓ so that (9) is met on the increasing patches of A, with ℓ constant on the flat patches of A, and satisfying (10) — with $A_\ell = A$ — for every left edge a_k of every patch for A. Standard arguments for differential equations ensure that ℓ is well defined, uniquely so up to $\ell(\underline{a})$. By Proposition 5, ℓ is incentive compatible. ■

Proof of Proposition 7. The proof will rely on the following lemmas:

Lemma 3. *Let $A \in \mathcal{A}$ and ℓ be the unique associated learning function as given by Proposition 6, with $\ell(\underline{a}) = \frac{A(\underline{a})}{c(\underline{a})-\lambda}$. Let dA and $d\ell$ be the Stieltjes measures associated with A and ℓ respectively. Then $d\ell$ is absolutely continuous with respect to dA and*

$$\frac{d\ell}{dA}(a) = \frac{1}{c(a) - \lambda}$$

is the Radon-Nikodym derivative of $d\ell$ with respect to dA.

Proof. First note that for any set $B \subset [\underline{a}, \bar{a}]$, $dA(B) = 0$ trivially implies $d\ell(B) = 0$ since A and ℓ are constant in the same intervals and strictly increasing in the same intervals. Hence $d\ell$ is absolutely continuous with respect to dA, and so there exists a Radon-Nikodym derivative between the two measures. Now let $[b, b') \subset [\underline{a}, \bar{a}]$. Then $[b, b')$ is made up of countably many intervals $[c, c') \in \mathcal{C}$ on which both A and ℓ are continuous and differentiable, along with at most countably many points of discontinuity $d \in \mathcal{D}$. It follows that

$$d\ell[b, b') = \ell(c) - \ell(b) = \sum_{(c,c') \in \mathcal{C}} \int_c^{c'} \ell'(x)dx + \sum_{d \in \mathcal{D}} \left[\ell(d) - \ell^{\uparrow}(d)\right]$$

$$= \sum_{(c,c') \in \mathcal{C}} \int_c^{c'} \frac{1}{c(x) - \lambda} A'(x)dx + \sum_{d \in \mathcal{D}} \frac{1}{c(d) - \lambda} \left(A(d) - A^{\uparrow}(d)\right) = \int_b^{b'} \frac{1}{c(x) - \lambda} dA(x)$$

where the third equality uses Proposition 5.

Since the set of intervals of the form $[b, b') \in [\underline{a}, \bar{a}]$ generate the Borel σ-algebra in $[\underline{a}, \bar{a}]$, we can conclude that $\frac{d\ell}{dA}(x) = \frac{1}{c(x)-\lambda}$ is the Radon-Nikodym derivative of $d\ell$ with respect to dA. ■

Lemma 4. ***Integration by Parts.*** *If P is an Q-integrable function on $[\underline{a}, \bar{a}]$, then Q is P-integrable on $[\underline{a}, \bar{a}]$ and*

$$\int_{\underline{a}}^{\bar{a}} P(x)dQ(x) = P(\underline{a}) \int_{\underline{a}}^{\bar{a}} dQ(x) + \int_{\underline{a}}^{\bar{a}} \int_{\underline{a}}^{x} dQ(y)dP(x)$$

Proof. If P is Q-integrable, then the standard integral by parts formula yields

$$\int_{\underline{a}}^{\bar{a}} P(x)dQ(x) = P(\bar{a})Q(\bar{a}) - P(\underline{a})Q(\underline{a}) - \int_{\underline{a}}^{\bar{a}} Q(x)dP(x) \tag{46}$$

Rearrange (46) to get:

$$\begin{aligned}\int_{\underline{a}}^{\bar{a}} P(x)dQ(x) &= \left[P(\underline{a}) + \int_{\underline{a}}^{\bar{a}} dP(x)\right] Q(\bar{a}) - P(\underline{a})Q(\underline{a}) - \int_{\underline{a}}^{\bar{a}} Q(x)dP(x) \\ &= P(\underline{a}) \int_{\underline{a}}^{\bar{a}} dQ(x) + \int_{\underline{a}}^{\bar{a}} \left(\int_{\underline{a}}^{\bar{a}} dQ(y(- \int_{\underline{a}}^{x} dQ(y) \right) dP(x) \\ &= P(\bar{a})Q(\bar{a}) - P(\underline{a})Q(\underline{a}) - \int_{\underline{a}}^{\bar{a}} Q(x)dP(x)\end{aligned}$$

■

Set $P = A$ and $Q = F_0$ in Lemma 4. Because F_0 is continuous and of bounded variation, the relevant integral is defined, and

$$\int_{\underline{a}}^{\bar{a}} A(a)dF_0(a) = A(\underline{a}) \int_{\underline{a}}^{\bar{a}} dF_0(a) + \int_{\underline{a}}^{\bar{a}} \int_{a}^{\bar{a}} dF_0(x)dA(a). \tag{47}$$

Next, setting $P = \ell$, and $dQ(x) = [\lambda - \sigma c(x)]dF_0(x)$ in Lemma 4, and noting again that Q is continuous and of bounded variation, we see that

$$\int_{\underline{a}}^{\bar{a}} [\lambda - \sigma c(a)]\ell(a)dF_0(a) = \ell(\underline{a}) \int_{\underline{a}}^{\bar{a}} [\lambda - \sigma c(a)]dF_0(a) + \int_{\underline{a}}^{\bar{a}} \int_{a}^{\bar{a}} [\lambda - \sigma c(x)]dF_0(x)d\ell(a). \tag{48}$$

Recall that $\ell(\underline{a}) = A(\underline{a})/[c(\underline{a}) - \lambda]$. Use this in (48), and invoke Lemma 3 to get:

$$\int_{\underline{a}}^{\bar{a}} [\lambda - \sigma c(a)]\ell(a)dF_0(a) = A(\underline{a}) \int_{\underline{a}}^{\bar{a}} \frac{\lambda - \sigma c(a)}{c(\underline{a}) - \lambda} dF_0(a) + \int_{\underline{a}}^{\bar{a}} \int_{a}^{\bar{a}} \frac{\lambda - \sigma c(x)}{c(a) - \lambda} dF_0(x)dA(a). \tag{49}$$

Combining (47) and (49),

$$
\begin{aligned}
\int_{\underline{a}}^{\bar{a}} [A(a) + (\lambda - \sigma c(a))\ell(a)]\, dF_0(a) &= A(\underline{a}) \int_{\underline{a}}^{\bar{a}} \frac{c(\underline{a}) - \sigma c(a)}{c(\underline{a}) - \lambda} dF_0(a) + \int_{\underline{a}}^{\bar{a}} \int_{a}^{\bar{a}} \frac{c(a) - \sigma c(x)}{c(a) - \lambda} dF_0(x) dA(a) \\
&= A(\underline{a})[1 - S(\underline{a})] + \int_{\underline{a}}^{\bar{a}} [1 - S(a)] dA(a), \qquad (50)
\end{aligned}
$$

where S is defined by

$$
S(a) = F_0(a) + \int_{a}^{\bar{a}} \frac{\sigma c(x) - \lambda}{c(a) - \lambda} dF_0(x). \qquad (51)
$$

Define $P = A$ and $Q(a) = 1 - S(a)$. S is continuous, $S(\underline{a})$ is finite and $S(\bar{a}) = 1$. We claim that S has bounded variation. Define $\Delta^+(x) \equiv \max\{\lambda - \sigma c(x), 0\}$ and $\Delta^-(x) \equiv -\min\{\lambda - \sigma c(x), 0\}$. Then, by (51):

$$
\begin{aligned}
S(a) &= F_0(a) + \int_{a}^{\bar{a}} \frac{\Delta^+(x)}{c(a) - \lambda} dF_0(x) - \int_{a}^{\bar{a}} \frac{\Delta^-(x)}{c(a) - \lambda} dF_0(x) \\
&= F_0(a) + \int_{\underline{a}}^{\bar{a}} \frac{\Delta^+(x)}{c(a) - \lambda} dF_0(x) - \int_{\underline{a}}^{a} \frac{\Delta^+(x)}{c(a) - \lambda} dF_0(x) - \int_{\underline{a}}^{\bar{a}} \frac{\Delta^-(x)}{c(a) - \lambda} dF_0(x) + \int_{\underline{a}}^{a} \frac{\Delta^-(x)}{c(a) - \lambda} dF_0(x).
\end{aligned}
$$

The first term on the right hand side of this equation is a cdf, nondecreasing in a. Consider each of the four integrals (without the sign that precedes them). Each integrand is a nonnegative-valued function (because $c(a) > \lambda$, and Δ^+ and Δ^- are each nonnegative), and each is nondecreasing in a (because $c(a)$ declines in a). Therefore, each integral is nondecreasing in a. It follows that S can be written as the sum/difference of five nondecreasing functions and consequently is of bounded variation. Therefore integration with respect to S is well-defined.

Now apply Lemma 4 yet again to (50) to obtain (12). ■

Proof of Theorem 3. Take any $A \in \mathcal{A}_R$. By Proposition 7, the ℓ function associated to A and with initial condition $\ell(\underline{a}) = \frac{A(\underline{a})}{c(\underline{a})-\lambda}$ has value $\int_{\underline{a}}^{\bar{a}} A(x)dS(x)$ for the S as defined. Now take another function ℓ' associated to A which satisfies initial condition $\ell'(\underline{a}) \in \left[0, \frac{A(\underline{a})}{c(\underline{a})-\lambda}\right)$. Then the value to the school of using ℓ' is

$$
\int_{\underline{a}}^{\bar{a}} A(x)dS(x) + \left(\frac{A(\underline{a})}{c(\underline{a}) - \lambda} - \ell'(\underline{a})\right) \int_{\underline{a}}^{\bar{a}} \lambda - \sigma c(x) dF_0(x)
$$

It is clear that, for any given A, if $\int_{\underline{a}}^{\bar{a}} \lambda - \sigma c(x) dF_0(x) > 0$, then this value is maximized by setting $\ell'(\underline{a}) = 0$. Otherwise, it is maximized by setting $\ell'(\underline{a}) = \frac{A(\underline{a})}{c(\underline{a})-\lambda}$. Hence, we know that for any given A, the maximal value to the school is given by

$$
\int_{\underline{a}}^{\bar{a}} A(x)dS(x) + \Theta A(\underline{a}) \qquad (52)
$$

where $\Theta = \max\left\{0, \int_{\underline{a}}^{\bar{a}} \frac{\lambda - \sigma c(x) dF_0(x))}{c(\underline{a}) - \lambda}\right\}$.

First, observe that, if $\Theta = 0$, then finding A^* which solves the categorization problem with R and S and setting ℓ to be associated to A^* and $\ell(\underline{a}) = \frac{A(\underline{a})}{c(\underline{a}) - \lambda}$ delivers the optimal learning function.

Now let's turn to the case when $\Theta > 0$. In Step 1, we fix $A(\underline{a})$ at a level in $\left[\underline{a}, \int_{\underline{a}}^{\bar{a}} x dR(x)\right]$, which is equivalent to fixing the first pooling interval to be $[\underline{a}, \alpha)$ for some $\alpha \in [\underline{a}, \bar{a}]$. We find A_α, the optimal categorization conditional on this first pooling interval by using the procedure described in Theorem 2. In Step 2, we use (52) to find the value of the best ℓ policy associated to each A_α and maximize over $\alpha \in [\underline{a}, \bar{a}]$ to find the overall optimal categorization. The optimal learning function is then associated to that optimal categorization and $\ell(\underline{a}) = 0$. ■

Proof of Proposition 8. Let Θ be as defined in Theorem 3. If $\Theta = 0$, then using the definition of S in (13) and applying Proposition 3 tells us that (16) is a sufficient condition for full pooling. If otherwise $\Theta > 0$, then by the definition of S and Proposition 3, we know that (16) is sufficient for full pooling to solve the problem for $\alpha = \underline{a}$ in Step 1 of Theorem 3. But when full pooling is the solution to the problem for $\alpha = \underline{a}$, then $\alpha = \underline{a}$ is the solution to Step 2, so we can conclude that (16) is sufficient for full pooling to be optimal. ■

REFERENCES

Alonso, R. and O. Camara (2016), “ ”Bayesian Persuasion with Heterogeneous Priors,” *Journal of Economic Theory* **165**, 672–706.

Boleslavsky, R. and C. Cotton (2015), “Grading Standards and Education Quality,” *American Economic Journal: Microeconomics* **7**, 248–279.

Boleslavsky, R. and K. Kim (2018), “Bayesian Persuasion and Moral Hazard,” SSRN 2913669.

Doval, L. and V. Skreta (2018), “Constrained Information Design: Toolkit,” mimeo. CalTech.

Dubey, P., and J. Geanakoplos (2010), “Grading Exams: 100, 99, 98,? or A, B, C?,” *Games and Economic Behavior* **69**, 72–94.

Dworczak, P. and G. Martini (2019), “The Simple Economics of Optimal Persuasion,” *Journal of Political Economy* **127**, 1993–2048.

Galperti, S. (2019), “Persuasion: The Art of Changing Worldviews,” *American Economic Review* **109**, 996–1031.

Guo, Y., and E. Shmaya (2019), "The Interval Structure of Optimal Disclosure," *Econometrica* **87**, 653–675.

Kamenica, E. and M. Gentzkow (2011), "Bayesian Persuasion," *American Economic Review* **101**, 2590–2615.

Kartik, N., Lee, F. and W. Suen (2019), "A Theorem on Bayesian Updating and Applications to Communication Games," *working paper*.

Kolotilin, A. (2018), "Optimal Information Disclosure: A Linear Programming Approach," *Theoretical Economics* **13**, 607–636.

Kolotilin, A. and H. Li (2019), "Relational Communication," *working paper*.

Lipnowski, E., Mathevet, L. and D. Wei (2019), "Attention Management," *AER Insights*, forthcoming.

Lizzeri, A. (1999), "Information Revelation and Certification Intermediaries," *RAND Journal of Economics* **30**, 214–231.

Moldovanu, B., Sela, A. and X. Shi (2007), "Contests for Status," *Journal of Political Economy* **115**, 338–363.

Ostrovsky, M. and M. Schwarz (2010), "Information Disclosure and Unraveling in Matching Markets," *American Economic Journal: Microeconomics* **2**, 34–63.

Popov, S. and D. Bernhardt (2013), "University Competition, Grading Standards, and Grade Inflation," *Economic Inquiry* **51**, 1764–1778.

Rayo, L. (2013), "Monopolistic Signal Provision," *The BE Journal of Theoretical Economics* **13**, 27–58.

Rayo, L., and I. Segal (2010), "Optimal Information Disclosure." *Journal of Political Economy* **118**, 949–987.

Rodina, D. and J. Farragout (2016), "Inducing Effort through Grades," mimeo.

Skreta, V. and L. Veldkamp (2009), "Ratings Shopping and Asset Complexity: A Theory of Ratings Inflation," *Journal of Monetary Economics* **56**, 678–695.