

Speeding, Tax Fraud, and Teaching to the Test

Edward P. Lazear

Hoover Institution
and
Graduate School of Business

Stanford University

Very Preliminary
January, 2004

This research was supported in part by CRESST.. Thanks to participants of the Stanford Lunch Seminar for useful comments. In addition, Gary Becker, Paul Oyer, Paul Romer and Kathryn Shaw Andy Skrypacz, Michael Spence, and Steven Tadelis were especially helpful in providing comments.

Abstract

Educators worry that high-stakes testing will induce teachers and their students to focus only on the test and ignore other, untested aspects of knowledge. Some counter that although true, knowing something is better than knowing nothing and many students would benefit even by learning the material that is to be tested. Using the metaphor of deterring drivers from speeding, it is shown that the optimal rules for high stakes testing depend on the costs of learning and of monitoring. For high cost learners, and when monitoring technology is inefficient, it is better to announce what will be tested. For efficient learners, de-emphasizing the test itself is the right strategy. This is analogous to telling drivers where the police are posted when police are few. At least there will be no speeding on those roads. When police are abundant or when the fine is high relative to the benefit from speeding, it is better to keep their locations secret, which results in obeying the law everywhere. Children who are high cost learners are less likely to learn all the material and therefore learn more when they are told what is on the exam.

High-stakes testing, where teachers, administrators, and/or students are punished for failure to pass a particular exam, has become an important policy tool. The “No Child Left Behind” program of the George W. Bush administration makes high-stakes testing a centerpiece of its approach to improving education, especially for the most disadvantaged. Proponents of high-stakes testing argue that testing encourages educators to take proper actions and that testing also identifies those programs that are failing.¹ But critics counter that high-stakes testing induce educators to teach to the test, which has the consequent effect of ignoring important areas of knowledge.² Almost every teacher is familiar with the question, “Will it be on the final?” The implication is that if it will not be, the student will not bother to learn it.

Which argument is correct? Is this an empirical question or one on which theory can shed light? The view taken here is that it is first necessary to define and set up the question in a rigorous way. After doing that, the answer becomes quite clear. In a nutshell, the main result is that to maximize the efficiency of learning, high-stakes testing should be used when learning and monitoring learning are very costly, but should not be used when learning and teaching monitoring is easy. This leads to interesting, but controversial policy conclusions.

The best way to focus the question, is to examine another problem that is formally

¹Identification is particularly important if, as Hanushek, Kain, and Rivkin find, teacher specific effects go a large way to explaining the performance of their students.

²See Koretz, et. al. (1991)

equivalent, namely that of deterring speeding.³ Suppose that the city has available to it a given number of police, who patrol the roads. Should the city announce the exact location of the police or simply allow drivers to guess? At first blush, the answer seems obvious. Of course, their locations should be kept secret. If the locations of the police are announced then motorists will obey the law only at those locations, but speed on at all other locations. But the answer is not obvious. If police locations are announced, at least speeding will be deterred in those places where police are posted. The total amount of speeding could actually be lower when locations are announced.

Tax fraud is virtually identical. The tax authority can announce the items to be audited, or just let taxpayers know that there will be random audits. In the absence of announcing specific items to be audited, taxpayers may cheat on all tax items, especially when there are few auditors and audits are unlikely. Instead, the authority can announce those items that will be audited with certainty and thereby deter cheating on those items, which is better than failing to deter any cheating.

Teaching to the test is analogous because the body of knowledge is like all of the roads. Announcing the items to be tested is like telling drivers which miles of road will be patrolled. If the test is not announced, but instead some random monitoring is done, students will have to decide whether to study a large amount or very little. When they would choose to study very little or nothing, announcing what is on the test may motivate them to learn at least those items. With the exception of definitions and some other formalities, the problems are the same.

Because the speeding model is the most straightforward and serves as the basic metaphor,

³There is a larger literature that begins with Becker (Crime and Punishment) on optimal incentives for enforcement of the law.

we begin by modeling it.

A Model of Speeding

There are Z miles of road. A driver can either speed or obey the speed limits. Suppose the extra utility that is derived from speeding is V per mile and that the fine for speeding, if caught, is K . There is a vast literature on optimal fines, but that is not the point of this example, so the fine is assumed to be given exogenously.⁴

Suppose that there are G police and that each policeman can patrol one mile of road. If police are distributed randomly along the road then on any given mile, the probability of being caught speeding is G/Z and the expected fine from speeding is

$$K G / Z .$$

Thus, if drivers do not know the location of the police, they will speed if

$$(1) \quad K G / Z < V$$

Since the cost and value of speeding on every mile is the same, if the driver chooses to speed on one mile, he speeds on all.

Now suppose that the location of the police along the roads is announced. Then drivers are certain to be caught if they speed on a patrolled section. As a result, no speeding occurs on the patrolled section as long as $V < K$, but speeding occurs on all non-patrolled roads because the drivers know that the probability of detection there is zero. As a result, if the location of the police

⁴In the teaching case analyzed below, the loss may be market determined, and then K is given exogenously to the student or teacher. As such, the model with exogenous fines is more appropriate for the main task of the analysis.

is announced, the law will be obeyed on G miles of road, and there will be speeding on the other $Z-G$ miles.

If locations are unannounced, there is either no speeding at all or always speeding. If locations are announced, then there is speeding on $Z-G$ miles, but not on G miles. Assume that it is desirable to deter speeding. Then it is better to announce locations of the police when

$$(2) \quad K G / Z < V < K.$$

If $K G / Z < V$, drivers would always speed were locations secret because the probability of detection is sufficiently low to make it worth the gamble. But announcing the locations deters speeding on G miles (since $V < K$) so this is the better outcome. If instead, $K G / Z > V$, the expected fine is sufficiently high to deter all speeding when locations are secret and this dominates revealing locations.

The intuition is simple. If police are few, drivers assume it very unlikely that they will be caught speeding and speed everywhere. Announcing locations of the drivers at least deters some speeding. If police are abundant, and the probability of being caught sufficiently high, no one will speed, but if locations of the police are revealed, drivers will speed on all roads except the G miles that are patrolled. With many police, it is better to keep their locations secret; with few police it is better to reveal their locations and at least deter speeding on the few roads that are patrolled.

The answer to whether locations should be revealed hinges on the costs of police. If the cost of enforcement is low or alternatively, if expenditures on enforcement is high, it is best to keep locations secret. When the costs are high or expenditures are low for budgetary reasons, it is better to reveal location.

After the analysis, the point seems obvious. But it is also somewhat counterintuitive. One might think that if there were many police, revealing their locations would be a good thing to do because most areas would be covered and so speeding would be deterred in most places. What is wrong with this logic is that the effect of adding police has an even greater payoff when the strategy is to keep them secret. In that case, adding police has a large and discontinuous effect. For G such that $K G/Z$ is just below V , the driver chooses to speed everywhere. For G such that $K G/Z$ is just above V , the driver chooses to speed nowhere. So in that narrow range, the marginal return to adding a policeman becomes very large. When police locations are kept secret, the marginal value is continuous and constant. Each additional policeman reduces speeding on one mile of road.

This logic implies that as long as police are costly, there is an optimal number of police. When police locations are secret, it is never optimal to have more police than

$$G = V Z / K,$$

which makes (1) hold with equality so that cheating is completely deterred.

Even when $G < V Z / K$, there is a hybrid scheme that improves on the one where patrolled miles are announced. Rather than giving specific locations of the police, it would be possible to declare that certain roads are subject to patrol with some probability so that

$$V = p K$$

where p is the probability that a given mile of road is patrolled. For some roads, it is zero, but for other roads, $p > 0$. For example, if $K = 2 V$, drivers could be told that there is one patrolman and that he is either on mile A or mile B, but nowhere else. Then drivers would be deterred from speeding on 2 miles of road per patrolman, rather than one.

Tax Fraud

The extension of the idea to tax fraud is straightforward. The tax authority can do random audits, examining taxpayers and items without advance notice or they can announce that all deductions of a particular kind will be audited. If they announce the items to be audited, taxpayers will report their expenditures honestly on the audited items. If they do not announce, then taxpayers will either cheat profusely or not cheat at all. If the cost of auditing is high or if there are very few auditors, it is better to announce the items that will be audited. Then, at least some taxes get paid honestly.. If the cost of auditing is low or if the expenditures on auditors is high, it is better to leave the identity of those to be audited and the items to be checked secret. Because auditing is sufficiently likely, taxpayers will be honest on all items.

The model is identical. V can be thought of due on each of the Z items. As such, it is the saving on taxes that results from cheating on one of Z reported items on the tax form. K is the fine associated with being caught, which includes repayment of the V dollars initially saved. Thus, $K > V$. G can be thought of as the number of items that can be audited (per return), given the number of tax auditors.

As before, when

$$(2) \quad K G / Z < V < K$$

filers will cheat on every item if monitoring is stochastic and will pay the penalty on those items on which they are caught. If the goal is to deter cheating, then a better system is to announce all of the items that will be audited and to deter cheating at least on those items that are audited with certainty.

If the goal is revenue collection, the argument does not hold. With stochastic monitoring,

individuals pay zero taxes initially, but are caught on G items (on average) and so pay $G K$ in total. With announced auditing rules, individuals pay taxes on the G announced items, which equals $G V$ in total. No fines are ever collected because the items on which the individuals cheat go undetected with certainty. Because $V < K$, revenues are higher in the stochastic monitoring regime than in the other regime.

The difference between the tax auditing problem and the speeding problem is that in speeding, the assumption is that the social cost of speeding is sufficiently high to swamp any distortions associated with reduced fine collection that might be part of an optimal tax structure. Here, if taxes are not collected through fines by the tax authority, the revenues must be raised in other ways, which may create other distortions. The goal taxing, at least in large part, is revenue collection.

Teaching to the Test

The lesson of the speeding example can be applied in a straightforward way to the issue of high-stakes testing. High stakes testing places the learning emphasis on announced items, and ignores others. In this sense, it is similar to the idea of announcing where the police are posted. The items on the exam receive special attention whereas untested items may be neglected by students and teachers. The speeding model can be applied to this problem in an almost direct fashion to obtain some insights. As above, the first result will be that high-stakes testing is best used when monitoring is costly or when expenditures on enforcement is low. If expenditures on enforcement

is high, then it is better to leave the testing regime more open.

Let us start by defining the knowledge base, which consists of n items. This is analogous to the z miles of road above. Suppose further that there are m questions on a high-stakes exam, analogous to the G policemen. First, let us simply think about revealing the questions that are on the exam versus keeping them secret. This does not appear to be the same question as whether high stakes tests should be used or not, but below it will be argued that one interpretation of the model is that the regime where questions are unannounced is like the world where students are not tested by any single test, but where they are monitored occasionally by less direct measures than the high-stakes tests. Further, high-stakes testing is as much about motivating teachers as students and the model applies to teachers as well. Initially, however, think of the student as making the choice about learning and let the teacher be a passive agent. That assumption will be altered below.

To be consistent with the speeding model, the return side is modeled as follows. If a student is asked a question to which he does not know the answer, he bears cost K in the form of lower earnings. Let us reinterpret V and K from the speeding model as follows: If the student does not learn the item, he does not have to bear cost V of learning the material. The student knows what is on the test, so opts to avoid learning an item when the extra utility from not learning, V , exceeds the cost of not learning, which is lost earnings, K . If the student knows what is on the test, he will choose to learn those items if and only if $V < K$. Since $K=0$ for items not on the test, he learns nothing that is not to be asked explicitly.

Now consider what happens when the student is told that testing is random. Let us think of m/n as measuring the probability that a student will be held accountable for any given item in the

body of knowledge with $0 \leq m \leq n$. This is meant to mimic the system without high stakes testing, where accountability is enforced only randomly and not necessarily through formal testing. But to make the comparison appropriate, it is necessary the similar intensities of monitoring across the two regimes are compared.⁵ Such monitoring could be on input or output. Students could be randomly questioned about material that they had learned. Teaching methods could be monitored on a random basis, as could teacher knowledge and curriculum. All of these are meant to be proxied by the stochastic monitoring intensity given by m/n . If $m=n$, then the monitoring intensity is such that all items in the knowledge set are monitored.⁶

Since costs are constant (relaxed below), if the student learns one item he learns all. The student will choose to learn an item and therefore every item when the expected cost of being caught unprepared exceeds the expected benefit of not studying. Thus, the student learns when

$$V < m/n K.$$

If this condition is reversed, the student will not learn at all.

As in the speeding model, high stakes testing produces more learning when, in the absence of revealing the specifics of the test, the individual would chose not to learn anything, but when the

⁵The teacher could even mimic the high-stakes test and announce the questions on it, which would make that one of a variety of possible test styles. But it is assumed that the student does not know that when learning the material. This is realistic in part because learning builds on other learning so a 7th grade student cannot know what kind of accountability will face him when he reaches 12th grade, in part because he does not know the identity of his 12th grade teacher.

⁶One technical difficulty is that the m that is associated with a given level of monitoring in a stochastic regime may not have the same cost to administer as asking m questions in a high-stakes test environment. Obviously, if costs are different, this will push the solution toward the lower cost alternative.

value of learning is sufficiently high that were questions announced, the individual would learn that material specifically. The condition for this to hold is

$$(3) \quad m/n K < V < K .$$

The left inequality implies that the student will learn nothing in the regime with stochastic monitoring, but will at least learn the m items when there is an announced, high stakes test.

The model produces some implications. First, notice that if $V < m/n K$, the student learns, even when monitoring is done only randomly. Then, more is learned by generalized, stochastic monitoring than by a high-stakes test where questions are known because the student learns only that material in the latter case. In order to prefer the high stakes test with well-known questions, it is necessary that (3) hold. To get some intuition, rewrite (3) as

$$(4) \quad m/n < V/K < 1$$

so the ratio of V to K must be less than 1 (or no learning occurs), but greater than m/n or it is better to engage in randomized monitoring.

Suppose that $V/K < 1$. If it is not, there is little to discuss because then learning is so costly relative to its value that learning does not occur under any circumstance. First, consider V/K and how it might differ by family background. Children from disadvantaged homes have higher costs of learning, V , and possibly lower returns to learning, K , than children from more affluent homes.

As a result, V/K is higher, meaning that disadvantaged students are less likely to master the entire set of knowledge. As a result, providing them with exactly the requirements for passing a test is more likely to result in greater learning for them.

The model captures exactly the intuition of both sides of the argument. Those who argue that announcing the exam questions will induce teaching to the test are correct if they are thinking about students who would be sufficiently motivated to learn all the material. These individuals are worried primarily about the relatively more able part of the distribution of learners. For them, $V/K < m/n$. Their costs of learning are sufficiently low relative to the returns to induce them to study the larger body of knowledge, even when holding them accountable for the material is random. But those who are the lower end of the ability distribution are in the opposite circumstance. For them, $V/K > m/n$. Their costs of learning are too high relative to the return and if they are not held accountable for a smaller subset of material, they will opt to learn nothing at all.

“No Child Left Behind” emphasizes high-stakes testing only for low performing schools. Although all schools are required to take the test, high performing schools are far away from the margin where anything is at stake. As such, the test has no monitoring incentive to those schools. If there is any monitoring incentive at all for upper quality schools, it is provided through more indirect stochastic methods. But failing schools are in the range where the high-stakes test matters. As a result, the NCLB system is essentially bifurcated, producing high-stakes testing for those who go to problem schools and stochastic monitoring (at best) for those who go to schools that are doing well. The model provides a rationale for this approach since the regime appropriate for low cost, low V students is exactly stochastic monitoring, whereas the one appropriate for high learning cost

children is likely to high-stakes testing.

A More Complete Model:

The constant cost of learning assumption above meant that students learned all or nothing. That assumption is sufficient to yield some insights, but it does not capture reality sufficiently and it is important to determine how things might change if we consider a smoother world. The answer is that things do not change very much.

Because the speeding model assumed a constant cost-constant benefit structure (the fine was always K and the value of speeding was always V), a driver chose to speed always or never. But it is unrealistic to assume that a student learns everything or nothing. More realistic is that there is a cost of learning, where a student who learns h items bears

$$\text{Cost} = C(h)$$

with $C', C'' > 0$. The convexity of the cost function reflects that the student has limited time and energy to put into learning and that some items are more difficult to learn than others.

Consider again the two regimes. In the first, the student is monitored stochastically. In the second, the student is required to pass a high stakes exam, the questions on which are known.

The maximization problem for the student is to choose the h questions to study to maximize earnings

$$(5) \quad \text{Earnings} = \text{Maximum Earnings} - K m [(n-h)/n] - C(h)$$

The student is monitored m times and on each time, he has $(n-h)/n$ chance of losing K .

The first order condition is

$$(6) \quad Km / n - C'(h) = 0$$

yielding an interior solution for h under the usual assumptions.⁷

Now consider the high stakes test regime where questions are known. The choice of high-stakes test over stochastic monitoring depends on the size of m , which interpreted as the number of questions on the exam or the intensity of monitoring in the stochastic regime. For now, the number of questions is taken to be exogenous, but that will be relaxed below.

The general result is this. If C'' is very low so that the marginal cost of learning rises slowly, then the stochastic monitoring approach is better. If C'' is large so that the marginal cost of learning rises sharply, then it is probably better to use a high stakes test. The issue revolves around the value of $C'(h)$ when $h=m$. A necessary and sufficient condition for stochastic monitoring to dominating high-stakes testing is that

$$(7) \quad C'(m) < K m/n .$$

The amount of learning that a student does under the high stakes test regime is

⁷Sufficient are that $C'(0) = 0$, and that $C'(h)$ goes to infinity as h goes to infinity. Note the second-order condition is $C'' > 0$, guaranteeing a minimum.

$$(8) \quad a. \quad h=m \quad \text{if } C'(m) \leq K$$

or

$$b. \quad h \text{ such that } C'(h)=K \quad \text{if } C'(m) > K$$

because the value of learning a question is K with certainty when it is on the test. If $C'(m) < K$, then the marginal cost of learning is below K even at m so all questions are learned. It would never pay to learn any more because no additional information is tested. Further, if the marginal cost of learning exceeds K before h equals m , then the student does not even learn all of the material that is on the test.

Note that in the regime with stochastic monitoring (eq. (6)), the student sets

$$C'(h) = K m/n$$

which cannot imply more learning than setting $C'(h)=K$ (eq. (7b)) because $m/n < 1$. Thus, the only way that the stochastic monitoring regime can result in more learning than the high stakes testing regime is if the student stops at m under high-stakes testing, but goes beyond that under stochastic monitoring. This can only happen if

$$C'(m) < K m/n.$$

Then the student will choose to set $h > m$ under stochastic monitoring, but only $h=m$ under high-stakes testing.

Once again, the spirit of this result is the same as the spirit of the result in the less realistic model, where complete or no learning occurs. When learning is cheap, stochastic monitoring dominates. When learning is expensive, high stakes testing is better.

Comparative Statics:

Using (6) and (8), it is straightforward to derive some comparative statics. In doing comparisons, it is necessary to ask first how the optimum changes within regime and second, whether the change implies a potential regime switch.

First consider a change in m , the amount of stochastic monitoring or the number of questions on the high stakes test. An increase in m is interpreted as a reduction in the cost of monitoring (modeled more explicitly below).

If monitoring is stochastic, then to see the effect of changing the intensity of monitoring on learning, use the implicit function theorem on (6) to obtain

$$\frac{\partial h}{\partial m} = \frac{-\partial / \partial m}{\partial / \partial h} = \frac{K/n}{C''} > 0$$

Increasing the intensity of monitoring raises the cost of ignorance because it is more likely that the student will be held accountable. Incentives to learn are thereby increased.

The story is somewhat more complicated because increasing m may involve a regime change. Recall that the condition for stochastic monitoring is that

$$C(m) < K m/n$$

An increase in m increases both sides of the inequality, and depending on the size of C'' , raising m could switch the optimal regime from stochastic monitoring to high stakes testing. Intuitively, raising m raises the amount learned under high stakes testing as well as under stochastic monitoring. A change in m could, conceivably, raise the amount learned under high stakes testing by more than

it raises the amount learned under stochastic monitoring. In either case, though, learning must increase. If there is a change from stochastic monitoring to high-stakes testing, this can only result in an even greater increase in the amount of learning that occurs with an increase in m .

Changes in K and n have even more transparent effects. Using (6) and the implicit function theorem, it is straightforward that increasing K unambiguously increases learning under stochastic monitoring and never results in a change in regime. Within the stochastic monitoring regime, it is clear from (6) that h is increasing in K . As the value of learning rises, students learn a greater proportion of the knowledge set. Furthermore, the condition for stochastic monitoring, (7), becomes more likely as K increases. If stochastic monitoring was optimal with the initial value of K , it is certainly optimal at higher levels of K .

Additionally, if the choice is to use high stakes testing, then the student has either already opted to learn all m items or has set $C'(h)=K$. In the latter case, where $h < m$, an increase in K increases the amount of learning and in the situation where $h=m$ even with the lower K , there is no change. Thus, in either regime, h is non-decreasing in K .

Again from (6), it is direct to show that opposite results pertain to n . If stochastic monitoring was initially chosen, then increasing n reduces the amount of learning because it reduces the probability that the student will be accountable for any given item and thereby reduces the return to studying. Furthermore, an increase in n could shift the choice of regime from stochastic monitoring, where $h > m$, to high stakes testing where $h \leq m$. Although this mitigates the reduction in learning that occurs with an increase in n , the amount learned necessarily decreases because stochastic monitoring is chosen only when the optimality condition results in $h > m$. Thus a regime

shift implies that h falls from $h > m$ to $h = m$ because (7) guarantees that $C'(m) < K$. This must be true because $C'(m) < Km/n_0$ where n_0 is the initial level of n in order to have chosen the stochastic regime initially. Note also that the amount of learning in under high-stakes testing does not depend on the number of items in the knowledge base. This is analogous to the amount of miles driven at the speed limit being unrelated to the total number of miles. If police are on G miles of road, those are the only ones on which drivers obey speed limits and the number of unpatrolled miles does not affect that. Students only learn (at most) m tested items, and increasing the number of potential items to learn that are not going to be tested has no effect on learning the tested ones.

Summarizing, increasing the value of learning unambiguously increases the amount of learning. Increasing the size of the knowledge base can only decrease or leave the same the amount of learning that occurs and may result in a change from stochastic monitoring to high-stakes testing. Reducing the cost of monitoring, modeled as increasing m , increases the number of items learned and may involve a change from stochastic monitoring to high-stakes testing.

Costly Testing and Heterogeneity:

If it were free to test, then one could simply set $m = n$, which would ensure the efficient amount learning in either the stochastic monitoring or high-stakes testing regime. In the stochastic learning context, the first order condition becomes

$$C'(h) = K \text{ if } h \leq m$$

and

$$h = m \text{ if } C'(m) < K .$$

These conditions imply efficiency when K is the social value of learning and C' is the marginal cost. In the high-stakes test regime, the student learns all m items as long as $C'(m) \leq K$ and $h < m$ items if $C'(m) > K$, in which case he sets h such that

$$C'(h) = K$$

just as with stochastic monitoring.

There are two issues that make the problem more interesting and more difficult for policymakers. The first is that testing and monitoring are not costless. The second is heterogeneity. Some students are better learners than others, so the optimal number of questions or monitoring for one student is not the same as the optimum number for another. If learning costs were observable, then it would be possible to adjust the amount of accountability to which each student is held so as to guarantee optimal learning for all. But acquiring such information is a tall order and using it in a way that was non-discriminatory might be even more difficult.

Implicit in the discussion to this point was that m was exogenous. Just as in the speeding problem, where the number of police was given and not subject to choice, in this application, it has been assumed that m , the amount of monitoring, is given. All that was to be determined was whether that monitoring should be done in a random way or with a high-stakes test. To consider the choice of m in a world of heterogeneous students, the assumption of a perfectly inelastic supply of test questions must be relaxed. The other extreme, used in this next section, is to assume instead that questions can be produced at cost t , allowing for a perfectly elastic supply at that price.

Now the choice becomes one of choosing m , recognizing that it is costly to do more monitoring. The problem can be analyzed formally.

The social value of learning is hK because each of the h items learned raises earning power by K . To allow for heterogeneity, rewrite the cost of learning h items as

$$\text{Cost} = C(h) / A$$

so that the social cost of learning is $C(h) / A$ for individual of ability type A . Further, let the cost per question asked by given by t .

Then the problem for society is to maximize the per-student expected social value or

$$(9) \quad \text{Max}_m \int [Kh - \frac{C(h)}{A}] f(A) dA - mt$$

The first-order condition is then⁸

$$(10) \quad \int [K \frac{\partial h}{\partial m} - \frac{C'(h)}{A} \frac{\partial h}{\partial m}] f(A) dA - t = 0$$

In the case of stochastic monitoring, h is determined by (6), which given the modified cost function says that

$$(11) \quad C'(h) / A = K m/n$$

⁸This assumes that A is such that the f.o.c. in (6) always gives an interior solution. If A were sufficiently high, it is possible that a corner solution of $h=N$ would be relevant, in which case, the expression in (10) would have branches.

To get a feel for the result, let us consider a specific example. Let $C(h) = h^2 / 2$. Then (11) implies that

$$h = AKm/n$$

In order to retain proper dimensionality, A is normalized such that AK has the same dimension as N . (This is simply to guarantee that h is expressed in units that are similar to n since they both refer to number of items in the knowledge set.) Substituting the solution for h into (10) yields

$$(12) \quad \begin{aligned} m &= \left(1 - \frac{tN}{AK^2}\right)N \\ &= \left(1 - \frac{\alpha N}{AK}\right)N \end{aligned}$$

where $\alpha = t/K$. If $\alpha = 0$, then $m = n$. When testing is costless, it is efficient to set m to n so that the student sets

$$C'(h)/A = K,$$

i.e., the marginal cost equal to the social value of learning. In general, $m < N$ because testing is costly. (Recall, that the normalization has AK in units of N and (12) only holds for $0 \leq m \leq N$.)

The primary result is that m is chosen to trade off the social value of learning with the cost from providing the incentive to learn. In general, it will not be optimal to choose m such that all the material is learned by all students.

The problem is messier in the high stakes testing regime. There, students learn $h \leq m$, but

since the actual questions are announced, as a practical matter many students might find it optimal to learn all m items on the exam. (It is much less likely that students will learn all N items in the knowledge set. As a result, the assumptions that imply interior solutions are more realistic in the stochastic monitoring problem.) Thus, corners are common and the problem in (10) must be rewritten to allow for $h=m$ for a significant fraction students.

In the high-stakes regime where questions are known, students choose h as

$$(13) \quad \begin{array}{ll} \frac{C'(h)}{A} = K & \text{if } \frac{C'(m)}{A} > K \\ \text{and} & \\ h = m & \text{if } \frac{C'(m)}{A} \leq K \end{array}$$

An interior solution (with $h < m$) is obtained for individuals who have sufficiently low ability defined as $A < A^*(m)$ where $A^*(m)$ is given by

$$(14) \quad A^*(m) = C'(m) / K$$

The equation for expected social value per student now has two branches: one for which $h=m$ and one for which $h < m$ is given by the interior f.o.c. in (13). Thus, the problem of choosing m in the high stakes test regime is

$$(15) \quad \text{Max}_m \int_{\frac{C'(m)}{K}}^{A_{\max}} \left[Km - \frac{C(m)}{A} \right] f(A) dA + \int_{A_{\min}}^{\frac{C'(m)}{K}} \left[Kh - \frac{C(h)}{A} \right] f(A) dA - mt$$

with first-order condition

$$(16) \quad \int_{\frac{C'(m)}{K}}^{A_{\max}} \left[K - \frac{C'(m)}{A} \right] f(A) dA - \left[Km - \frac{KC(m)}{C'(m)} \right] f\left(\frac{C'(m)}{K}\right) +$$

$$\int_{A_{\min}}^{\frac{C'(m)}{K}} \left[K \frac{\partial h}{\partial m} - \frac{C'(h)}{A} \frac{\partial h}{\partial m} \right] f(A) dA + \left[Kh - \frac{KC(h)}{C'(m)} \right] f\left(\frac{C'(m)}{K}\right) - t = 0$$

Even in simple cases, the expression does not provide much intuition. The approach, numerically, would be to plug in the functions, calculate the optimum values of m for each regime using (10) and (16), and then use the optimal m values to evaluate (9) and (15). The regime chosen is the one that yields the highest social value.

Separating Teacher and Student Incentives:

The discussion has been put in terms of motivating students, but most of the thought behind specific programs like “No Child Left Behind” is that it is the teacher, not the student who needs motivating. At the most abstract level, the model as set up can be interpreted to refer to teachers instead of students.

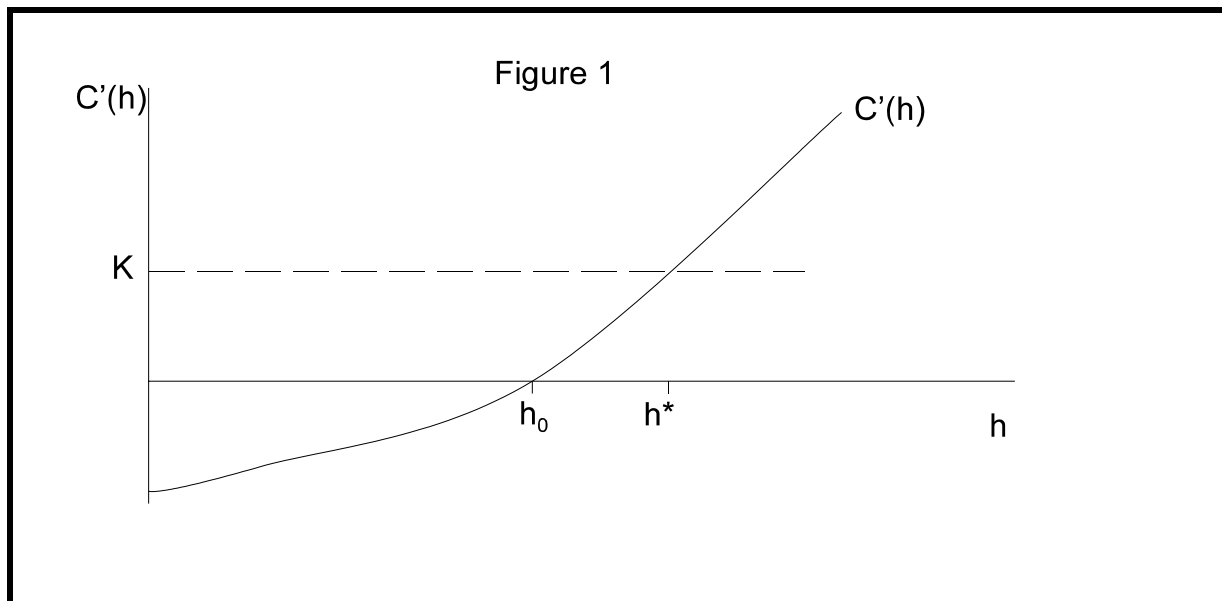
Suppose that teachers have full control over what is learned by the student. Interpret $C(h)$ as the teacher’s cost of teaching the student h items of knowledge. Let K be the penalty associated with her student failing to answer a question correctly in the high stakes environment or as the penalty that the teacher faces if the student is detected to be ignorant of an item of knowledge in the regime of stochastic monitoring. Then all of the above analysis holds exactly as written and nothing is changed.

The problem of interest, though, is how are teachers motivated. Many would argue that the current system of stochastic monitoring does not motivate teachers at all. Teachers are motivated by intrinsic considerations only and intrinsic motivation is insufficient to induce some teachers to do the right thing, either because they are lazy or lack the ability to do so. Again, the issue is one of heterogeneity as well as motivation, but let us consider the incentive issue in a world of homogeneous teachers first.

Intrinsic motivation might be thought to serve as the main motivator for tenured teachers whose salaries are fixed and jobs are secure and virtually independent of performance. Intrinsic motivation is best modeled by assuming that $C'(h) < 0$ for $h < h_0$ (on a per student basis). Even if teachers received no compensation for the amount of knowledge their students acquired, they would

still choose to provide h_0 of knowledge to each student. If compensation is positive, then in the stochastic monitoring regime, the first-order condition in (6) still holds because optimality always occurs for $C' > 0$.

Things are slightly more complicated for high-stakes testing. Before, the value of learning



(or teaching) additional items once $h=m$ was zero. Now, however, it is possible that a teacher might prefer to teach $h>m$ items because of her intrinsic motivation. Consider figure 1.

Suppose that in the high stakes test regime, $m < h_0$. Without intrinsic motivation, the teacher would set $h=m$ because there would be no reason to exceed m . The value of teaching another item would be zero because K is only relevant up to m . But with intrinsic motivation, the teacher still continues to teach until h_0 units are learned. For $h_0 < m < h^*$, the teacher would teach exactly m units

because she is already beyond the level of h provided for by intrinsic motivation and there is no direct value to teaching $h > m$. For $m > h^*$, the teacher provides h^* of knowledge where h^* is the solution to (8b), as above.

It remains true, however, that because $K > K m/n$, a necessary and sufficient condition to obtain more teaching under stochastic monitoring than under high-stakes testing is that

$$C'(m) < K m/n$$

as in (7). Because $K m/n > 0$, the case of $m < h_0$ is dominated by stochastic monitoring and this case also satisfies the condition in (7). For $m > h_0$, the cases are identical to those described above in (6) and (8) so nothing is changed.

What is different, though, is that (7) is more likely to hold when teachers are intrinsically motivated. Intrinsically motivated teachers have lower values of C' for any given h than otherwise identical teachers (by definition of intrinsic motivation). As such, the condition

$$C'(m) < K m/n$$

is more likely to hold for intrinsically motivated teachers.

This analysis yields the result that when teachers are intrinsically motivated, a stochastic monitoring regime is more likely to dominate the high-stakes testing regime than when teachers are not intrinsically motivated. Expected effort is higher in both cases (because there will be cases where $h_0 > m$ in high stakes testing), but the difference is more relevant for stochastic monitoring than for intrinsic motivation. Put differently, stochastic monitoring is relatively less effective for less motivated teachers. This is true even when the size of the penalty and monitoring intensity is the same in both regimes. So No Child Left Behind, which provides incentives for poorly

performing teachers and schools, could well be a step in the direction of efficiency.

Other issues with teachers and students involve team problems. Because both have an incentive to free ride on the other's effort, the standard result that effort of each party falls short of the optimum holds. But there is little about the student - teacher team that distinguishes it from other partnership problems, which have been analyzed.⁹

Adjusting the Cost of Failing:

The penalty from missing a question, K , has been assumed exogenous. When students are the agents, this seems the right assumption because K is a market determined parameter, given by the amount that employers reduce wages for every question missed on the exam. Wage reductions reflect firms' best estimate of lower ability associated with failing to know the answers to questions on the exam. As a result, K is not a policy variable in this context.

When teachers are the agents, and students are assumed to be totally passive in the process, then K is a policy variable because it is then interpreted as the amount by which teachers are penalized for each question that her student fails to answer correctly. By increasing K , performance is enhanced. Of course, raising the penalty also means that the base wage must be adjusted upward in order to induce individuals to enter teaching.

Much of the "high-stakes test" debate revolves around raising K so as to penalize teachers more. An increase in K places more emphasis on extrinsic and less on intrinsic motivation. As long as teachers are risk averse and test results have a stochastic component, raising K imposes more risk

⁹See Kandel and Lazear (1991).

on teachers. One way to mitigate that risk is to provide the teachers with information as to which items will be on the test so that the teacher can reduce the likelihood that the student will study untested material. As before, this has a cost by reducing the total amount of learning, but it has the benefit that it reduces risk and improves incentives to teach low ability students for whom the cost to the teachers is highest. The details of the optimal mechanism are left for future work.

Test Design and Learning Incentives:

Stochastic monitoring looks at one item at a time. In high-stakes testing, an exam is given at one point, but it is still true (at least as modeled), that the payoff is linear in questions answered correctly. There is no passing or failing of exams. In this section, we consider whether it is better to treat each question independently or whether to have an exam where there is a requirement that some number of questions be answered correctly in order to pass.

The theory is that of sampling with replacement. An item can be tested twice and some items never on any series requirements of explicit high-stakes exam or an implicit monitoring schemes. Suppose that a student must get s answers correct out of m questions where as before, there are n items in the knowledge base. With high-stakes testing, where questions are announced, little changes. The only difference is that instead of learning a maximum of m items, the student learns a maximum of s items because there is no value to learning the additional $m-s$ answers. A pass is a pass. A passing grade on the test is worth

$$sK$$

in this regime because every employer knows that the student knows exactly s items and would be

assumed to not know the other $n-s$ items.

In the stochastic monitoring world, the problem is somewhat more complicated. First, probability theory yields the result that the probability of getting exactly s correct on an exam of m questions is

$$(17) \quad \text{prob}(s \text{ correct} \mid m \text{ questions}) = \frac{\binom{m}{s} h^s (n-h)^{m-s}}{n^m}$$

The probability of a pass is then

$$(18) \quad \text{prob}(\text{pass}) = \sum_{i=s}^m \text{prob}(i \text{ correct} \mid m \text{ questions})$$

In the special case where 100% correct is required to pass, (18) becomes

$$(19) \quad \text{prob}(\text{pass}) = \left(\frac{h}{n}\right)^m$$

The maximization problem for the student in this regime is in general

$$\begin{aligned} & \underset{h}{\text{Max}} \text{ Maximum Earnings} - Km(\text{prob}(\text{fail})) - C(h) \\ & \text{or} \\ (20) \quad & \underset{h}{\text{Max}} \text{ Maximum Earnings} - Km(1 - \text{prob}(\text{pass})) - C(h) \end{aligned}$$

with first-order condition

$$(21) \quad Km \frac{\partial \text{Prob}(\text{pass})}{\partial h} - C'(h) = 0$$

In the case where all m must be answered correctly, (21) becomes

$$(22) \quad \frac{Km^2}{n} \left(\frac{h}{n}\right)^{m-1} - C'(h) = 0$$

Comparing (21) to (6), a higher h is chosen in the pass/fail environment than the one-at-a-time stochastic monitoring environment iff

$$\frac{1}{n} < \frac{\partial \text{prob}(\text{pass})}{\partial h}$$

At this stage, few analytic results have been obtained, but for a particular case, it is clear that the pass/fail method can never dominate.

Let $C(h)$ be a polynomial of the form

$$C(h) = h^a/a$$

Then

$$h = \left(\frac{Km^2}{n^m} \right)^{\frac{1}{a-m}}$$

Since $M(\text{prob pass}) / M = (h/n)^{m-1} / n$, substitution yields

$$(23) \quad \frac{\partial \text{prob}(\text{pass})}{\partial h} = \frac{m}{n} \left[\frac{(Km^2 n^{-m})^{\frac{1}{a-m}}}{n} \right]^{m-1}$$

When $m=1$, (23) becomes

$$\frac{\partial \text{prob}(\text{pass})}{\partial h} = \frac{1}{n}$$

It is also true that when $m=1$

$$\frac{\partial^2 \text{prob}(\text{pass})}{\partial h \partial m} = \frac{\ln \left[\frac{(K/n)^{\frac{1}{a-1}}}{n} \right]^{m-1} + 1}{n}$$

which is negative for combinations of K , n , m , and a that yield interior solutions. As a result, in the case of a polynomial, forcing a student to take one large (m at a time) stochastically designed test cannot dominate having m smaller tests or monitoring experiences. The “pass the entire exam” approach is identical to the “each question on its own” approach only when $m=1$ (so that they are identical). For $m>1$, the each-question-on-its-own approach dominates because

$$\frac{\partial \text{prob}(\text{pass})}{\partial h} < \frac{1}{n}$$

The proposition appears to be general, but this is only for the case where $s=m$. A conjecture is that it remains true for all cost functions that satisfy the necessary conditions for interior solutions and true even when $s < m$, so that the student is not required to get all right to pass the exam. This is left for future work.

Hybrid Schemes and Pop-Quizzes:

Suppose that

$$(24) \quad Km/n < C'(m) < K.$$

Then, by (7) the high-stakes test dominates, but there are wasted questions in the sense that the student learns no more than m items, but the constraint is slack. In the same way that miles were added to ranges over which police were posted, the same can be done here. Although all n items may not be target questions, the school could announce that some subset, q , of the n items are possible questions such that

$$C'(q) = Kp$$

where p now is interpreted as the probability that any one question will be asked. Then the constraint is tight. Since (24) holds, p is greater than m/n , but less than 1.

In this context, the high stakes test consists of naming a subset of the n items in the knowledge set, consisting of q items such that the probability that any one of the q items is on the exam is p . Because the q items are identified in advance, the number of items in the knowledge set, n , is irrelevant. All $n-q$ of them that are not in the subset q are ruled out. Thus, the student is told that there are q items to study and that m of them will be on the exam. (Take m as given exogenously.)

The probability that any one of the q items is on the exam is then

$$m/q$$

so q is chosen to make¹⁰

$$C'(q) = K q / m .$$

A pop-quiz is one type of hybrid scheme. Students are told that occasionally, there will be a test on well-specified items, like last night's reading assignment. They are not told when the quiz will occur. This is like telling them that q items are subject to testing (e.g., the material in the reading), but testing only m of it by reducing the probability of a test below one. This scheme increases the effectiveness of questioning.

Another interpretation of a hybrid scheme is one where a specified, high-stakes test is used, but it is coupled with stochastic monitoring, like the principal visiting the classroom to observe the input side of the process. Hybrid schemes can never be worse, of course, because a special case of the hybrid scheme is one that places all weight on one scheme. But there are two practical difficulties with hybrid schemes that choose interior solutions. First, it is not costless to observe input so the tradeoff is between more high stakes testing and some classroom observation or other

¹⁰ $C''(h)$ must be sufficiently high to guarantee an interior solution. For example, if $C=h^a/a$, then a necessary condition for an interior solution is that $a>2$. This requirement is not particularly stringent. Because of limited time, the slope of the $C'(h)$ function must eventually go to infinity. When a student has spent all of his time and effort learning, he has no ability to increase h beyond some point.

stochastic monitoring. It has already been shown that in some cases, high-stakes testing provides better incentives than stochastic monitoring so the tradeoff might sometimes favor corners. Second, the monitor of the teacher is the school principal and unfortunately, the principal is not the principal, but merely another agent. In order to determine whether the principal is evaluating his teachers properly, it is necessary to have some outside check on his evaluations. The natural check is an objective, high-stakes test. Any additional subjective evaluation merely pushes the question back one level.

Measured and Unmeasured Aspects of Learning:

One problem with high-stakes compensation of any form is that it induces individuals to focus on measured aspects and ignore unmeasured ones. This comes up in the context of paying piece rates, where quantity is cheaper to measure than quality, and piece rates induce workers to produce too many low quality items. This is sometimes referred to as the “multi-task” problem.¹¹ In the context of teaching, this might manifest itself as a focus on learning facts that are easily tested, but ignoring deeper more conceptual issues that are more difficult to assess.

There is no doubt that focusing on one type of education leaves other types untested, but that

¹¹See Lazear (1986) where the two dimensions of output are quantity and quality, Holmstrom and Milgrom (1991), where the two dimensions are attributes of output, one of which is more easily measured, and Baker (1992), where the dimensions consists of effort in different states of the world.

The problem here, technically, can be regarded as one of multi-tasking, because each item of knowledge is distinct and separately observable. But the key results of the quantity / quality, or multiple outputs is that not only are the outputs inherently different, but some are easier to observe than others (e.g., quantity v. quality).

issue is probably secondary in this context for a variety of reasons.

First of all, the problem that critics of high stakes testing worry about is not items that cannot be tested, but items that are simply ignored. For example, Daniel Koretz makes the point that a particular test always concentrates on regular polygon and never tests knowledge of irregular polygon. It is not more expensive to test knowledge of the latter, it is simply the case that one cannot test everything because testing is costly and testing patterns come to be known, so the tested items are learned and the untested items are not. The same is true with respect to evidence on different tests. When a group of students are shifted from one test, say, SATs, to another, say ACTs, they initially perform worse (in percentile terms) on the new test than they did on the old. Over time, average scores on the new test rise. When students are given the former test, they perform worse on it than they do on the new test and than they did before the switch.¹² Both tests are the same in that they test the same type of material, but different specific components of it. This issue here is not that some aspects are easier to measure than others, which is the emphasis of the multi-task literature, it is that some items are chosen for testing by one exam and ignored on the other exam.

Second and related, testing is quite sophisticated and advanced and abstract topics are tested all the time. Even college board exams have open form questions that test for creativity and writing ability. While grading this part might be somewhat more expensive than grading other parts, computerized grading of essay exams has made this distinction much less important. Indeed, at the graduate level, we teach very abstract concepts with relatively primitive tests, but most believe that

¹²Again, see Koretz, et. al. (1991).

our tests give us a good indication of student performance, and certainly of relative position within the class.

Third, for most students, especially at the K-12 level, creativity and other less easily tested items are not the key issue. Most of the discussion revolves around basic verbal and mathematical literacy, both of which are easily tested. Creativity and other difficult to measure components are important, but for a small part of the population and that group is in no danger of failing the basic tests anyway. Indeed, one result of the paper is that stochastic monitoring is more efficient for high ability students and high stakes testing is more efficient for low ability students.

For these reasons, this analysis has assumed that each of the n items in the knowledge base are perfect substitutes for one another in both production and in testing. Although not literally true, this is likely to be a good approximation for the issue that is central to the policy debate.

What is a “Good” Test?

One common view is a good test is one that is not so predictable that students essentially know what is on the exam. It would be possible to create an exam that randomized, avoiding the type of problems illustrated by the example of testing regular polyhedrons, but never testing irregular polyhedrons.

This view is incorrect. Although it may be optimal to construct a test that draws from a larger body of knowledge, the main theorem of this paper is that sometimes it pays to restrict the relevant required material to a specified, subset of the entire knowledge set. A “good” test when students have very high costs of learning is a test that announces the questions and sticks to them. Under

those circumstances, students at least learn the material that is on the test. The alternative test, which choose questions from a broader base of knowledge, results in no learning or very little learning.

For low cost learners, the reverse is true. A test that draws from the entire or a larger knowledge base is a better test because it encourages more learning than one that is well-specified and announced. For these students, a “good” test is one that is not completely predictable, because it provides more incentives to learn.

Thus, the definition of what is a good test is just another interpretation of the result of this paper. Corners, where all questions that will be on the exam are announced, are sometimes optimal and this happens when learning (or monitoring) costs are high. Randomized question choice corresponds to what has been called stochastic monitoring throughout.

Conclusion

Speeding, tax fraud, and teaching to the test are all symptoms of the same kind of incentive structure. Individuals become aware of the rules, obey them within a narrow range, and disregard them everywhere else.

The analysis in this paper has shown that providing high-stakes and well-defined requirements dominates stochastic incentives for individuals for whom compliance costs are high. In the context of education, this means that predictable high-stakes tests are best used for high cost learners or low ability types and stochastic monitoring, where students are not informed in exact terms what will be required of them, provides better incentives for low cost learners or high ability types.

Additional results are provided.

1. If teachers have low degrees of intrinsic motivation, then well-defined high stakes tests are best, but for teachers with high intrinsic motivation, a more randomized accountability system is efficient.

2. Smaller lower stakes tests may under certain circumstances provide better incentives than large, high stakes test. There are two reasons. First, better incentives are provided for a given number of questions by breaking up the test, i.e., giving some credit for getting some, if not all, correct. Second, pop-quizzes and other low-stakes, but somewhat randomized testing methods, may provided better incentives than either extreme of totally stochastic monitoring or completely specified tests.

3. A “good test” is a well defined concept once incentives are considered. Good tests are not necessarily those that draw evenly from the knowledge base, or even from the important knowledge base. Sometimes, especially for high cost learners or failing teachers, tests that are predictable are best at providing incentives to learn. For high ability students or successful teachers, somewhat more unpredictable tests are best.