Speeding, Tax Fraud, and Teaching to the Test

Edward P. Lazear

Hoover Institution
and
Graduate School of Business

Stanford University

September, 2004

This research was supported by CRESST. I am indebted to Edward Glaeser for some of the derivations contained in the paper. In addition, George Akerlof, Gary Becker, Thomas Dohmen, Richard Freeman, Eric Hanushek, Caroline Hoxby, Larry Katz, 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 this may be 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 police 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. The same logic also implies that tests should be clearly defined for younger children, but more amorphous for more advanced students.

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 induces 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 on the final, the student will not bother to learn it.

[1] Identification is particularly important if, as Rivkin, Hanushek, and Kain (2001) find, teacher specific effects go a long way in explaining the performance of their students.

[2] See Koretz, et. al. (1991) discussed in more detail below.

Hoffman, Assaf and Paris (2001) report on results from Texas Assessment of Academic Skills testing. Using a sample of 200 respondents, they suggest that the Texas exam has negative impacts on the curriculum and on its instructional effectiveness, where 8 to 10 hours per week on test preparation is typically required of teachers (by their principals) and the curriculum is planned around the test subjects. They also argue that teaching to the test raises test scores without changing underlying knowledge.

Jones, et. al, (1999) study data from North Carolina and conclude from a survey of 236 participants that the high-stakes test induced two-thirds of teachers to spend more time on reading and writing and 56% of teachers reported spending more time on math. They also claim that students spend more than 20% of instructional time practicing for end-of-grade tests and a significant fraction report a reduction in students' love of learning.

In an early study, Meisels (1989) outlines some of the pitfalls of high-stakes testing and suggests adverse effects of the Gesell School Rediness Test and of Georgia's use of the CAT.

Which argument is correct? The main result of the following analysis is that to maximize the efficiency of learning, high-stakes, predictable testing should be used when learning and monitoring learning are very costly, but should not be used when learning and monitoring are easy.

The best way to focus the question is to examine another problem that is formally 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, and will speed at all other locations. But the answer is not obvious. If police are very few and their locations are unknown, drivers might decide to speed everywhere. If police locations are announced, there is a better chance that speeding will be deterred at least 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 auditedor 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 likely 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.

[3] Beginning with Becker (1968), there is a large amount of literature on optimal incentives for enforcement of the law.

Announcing the items to be tested is like telling drivers which miles of road will be patrolled. If the test questions are 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, we begin by modeling it.

A Model of Speeding and Tax Fraud

Deterring 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

[4] In 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.

(1) $K G / Z < V$.

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. A more general approach allows for some miles to be subject to patrol with some probability and others with some different probability, but to get the basic intuition, let us start with the more extreme version of the model. If roads are either patrolled or not, 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. 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, depending on whether the expected fine, KG/Z, exceeds or falls short of the utility value of speeding, V. But when locations are announced, there is speeding on Z-G miles, but not on G miles as long as $K>V$. Assume that it is desirable to deter speeding for all drivers in all situations. Then it is better to announce locations of the police when

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

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 speeding 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 police strengthens incentives on patrolled roads and at least deters speeding at those locations. If police are abundant and the probability of being caught sufficiently high, no one will speed. With many police, revealing their location induces drivers to speed on all roads except the G miles that are patrolled. So when there are 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.

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

$$G= V\,Z\,/\,K,$$

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

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. So, when the cost of auditing is high perhaps because there are very few auditors, announcing the items that will be audited results in less cheating. Announcing audited items ensures that at least some taxes get paid honestly. However, were the cost of auditing low and auditors abundant, then keeping audited items secret would result in more taxes being paid. Because auditing is sufficiently likely, taxpayers are honest on all items.

The model is identical. V can be thought of as tax 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, $K>V$. Redefine G as the number of items that can be audited (per return), given the number of tax auditors.

As before, when

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

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.

When G is high, there is no conflict between revenue collection and deterrence. Then, audit rules are not announced, no one cheats and ZV is collected. If audit rules were announced, only GV would be collected and $ZV>GV$ because $Z>G$.

When G is low (auditors are very costly), there is a conflict between deterrence and revenue collection. If G is low, more fraud is deterred by announcing the audit rules than by keeping them secret, but more revenue is collected by keeping rules secret. When G is too low to deter cheating if rules are unannounced, individuals cheat on all items, paying zero taxes, 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, and revenues are G V. No fines are ever collected because the items on which the individuals cheat go undetected with certainty. Because V<K, revenues are highest in the stochastic monitoring regime, even though no cheating is deterred.

Keeping the rules secret induces everyone to cheat, which is like setting a trap for cheaters. Entrapment can be useful for revenue collection because "tricking" people into cheating results in fine collection, which brings more money into the treasury than the paying of taxes without fines. But note that this structure is entrapment and trickery only in an ex post sense. Taxpayers know ex ante that they may be caught and weigh the probability and fine appropriately. They break the law consciously, weighing the risks. There is no fraud on the government's part nor is there an attempt to coerce any given taxpayer into taking an illegal action. Audit probabilities and fines are known in advance.

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[5]. 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 of 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

[5] Some might argue that speeding fines are part of an optimal tax structure. For example, some very small towns set excessively low speed limits to induce speeding so that they can collect revenue from out-of-town motorists.

high-stakes testing. High-stakes testing as a practical matter places the learning and teaching emphasis on items that are expected to be on the exam. 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 is 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. Second, high-stakes testing with well-defined exam questions is best used when the distribution is weighted toward high cost learners.[6]

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. Should the exam questions be announced or not? A more direct way to put the issue is "What comprises a good high-stakes test?" Should it be a test where questions are well-defined and known in advance, or should it be a test where questions are drawn randomly from a larger body of knowledge? Most would say the latter. It will be argued that the former rule is appropriate in some circumstances.

Further, as a policy issue, testing is as much about motivating teachers as it is about motivating 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

[6] The emphasis here is on the incentive aspect of testing. Another role of testing is to provide information to help in modifying the curriculum. To deal with this component of testing, a dynamic model is required.

will be altered below.

To be consistent with the speeding model, the return side is modeled as follows. Think of the test score as an observable signal to employers, or more accurately, to future schools which the student might attend. 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, most directly reflected as reduced probability of admission into a desirable college. The SAT exam is a high-stakes test with exactly that effect. The "fine," K, is taken to be exogenous, but a richer model would allow K to be the solution of an inference problem that colleges or employers make about the individual's ability based on the answers to the exam. Below, (see the section titled "*Inference*"), endogeneity of K is investigated in more detail. In this section, exogeneity of K is assumed for convenience.

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 he 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 $0\leq m\leq n$. Initially, suppose that V is the same on all items in the knowledge base and across all students. 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, all students learn everything when

$$V < m/n\ K.$$

If this condition is reversed, no student learns anything.

As in the speeding model, high-stakes well-defined 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 value of learning is sufficiently high that were questions announced, the individual would learn that material specifically. The condition for this to hold is

$$m/n\ K < V < K .$$

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

It is now necessary to model social costs explicitly and drop the assumption that V is identical for all people and for items in the knowledge set.

Allow there to be a distribution of V that reflects the cost of learning on any given item by any given person. Let that distribution be written J(V) with corresponding density j(V). The unit of analysis is a person-item so that V can vary for a given individual because some items are more difficult to learn than others. Also, V can vary across people because some people learn more easily than others. Then j(V) is the density across all items potentially learnable by all students. Note that a given student might learn some items and not others and some students might learn everything always and others nothing, depending on the distribution of V across items and people. Further note that the assumption is that all items and students characterized by the distribution J(V) are observationally identical. If items or people are observably different, then separate distributions must be written to characterize each. Finally, note that V is assumed to be independent of whether learning of other items occurs. For the sake of simplicity, such complications are ignored.

Suppose that there is some expected penalty, X. A given student learns an item if and only if V>X. Let the social value of learning be given by γ. Items for which V<X are learned. Those for which V>X are not learned. Thus, the social damage associated with any expected penalty X is

(3) $$S(X) = \int_X^\infty (\gamma - V)\, j(V)\, dV$$

Also note that

(4) $S'(X) = (X-\gamma)\, j(X)$

and that

(5) $SO(X) = jN(X)(X-\gamma) + j(X)$

which will be useful later. In what follows, optimal solutions are found in the more general analysis where a rich structure of strategies is considered.

From (4) and (5), it is clear that social damage is minimized when X=γ. Setting the expected fine equal to the social cost of the infraction induces the appropriate behavior.

Continuous Choice and Interior Solutions

In the simple speeding model, the enforcement choice was between two alternatives. Either police were stationed randomly over all existing roads or there was a section of road on which every mile was patrolled. A more general formulation allows for some miles to be subject to patrol and others not. In the speeding framework, given that there are G police, there could be that some proportion of all roads, q, that are patrolled and some proportion, 1-q, that are not patrolled. The California Highway Patrol (CHP) uses exactly this strategy. For example on the July 4 weekend of 2004, the CHP announced on all TV news stations that the 250 miles of Interstate 80 from San Francisco to the Nevada state line was being singled out to check for intoxicated drivers.

Similarly, certain items in the knowledge set can be made eligible for testing and others can be declared off limits. Let qn of the items in the knowledge base be subject to testing. When q=1, all items are fair game. When q=m/n, (m/n)n or m items are available for testing. Since there are m questions, each item subject to testing is identified.[7] Then, on eligible items, the probability of any given item being tested is one and it is zero for all ineligible items.

The problem then is to choose q so as to minimize social damage. It is shown in this section that corner cases are possible where the solutions are q= m/n (reveal the questions) or q=1 (do not reveal anything).

Recall that S(X) is the social damage on a given item when the extended penalty is X. Then, expected damage as a function of q is given by

[7] It is assumed that the same item is never tested twice.

(6) Full social damage / FSD(q) = n [q S(X) + (1-q) S(0)]

On the qn eligible items, damage is S(X) and on the (1-q) n ineligible items, the damage is S(0). The expected penalty on the eligible items is

$$X = K\frac{m}{q\,n}$$

so (6) can be written as

(7) FSD(q) = n [q S($K\frac{m}{q\,n}$) + (1-q) S(0)]

Differentiate with respect to q to obtain

(8) $$\frac{\partial}{\partial q} = n\left[\,S(K\frac{m}{qn}) - S(0) - K\frac{m}{qn}S'(K\frac{m}{qn})\right]$$

In order for it to be optimal to state exactly where the police are located, q must be equal to m/n. This happens if M/Mq in (8) is positive at q=m/n. Then, increasing q will only increase social damage so the corner solution is best.[8] The requirement is that

$$S(0) - S(K) < -KS'(K) \ .$$

[8] An additional requirement is the regularity condition that if M/Mq becomes negative, S(K)>S(KG / qZ).

A sufficient condition for this to hold is that S(X) is concave over the range of X from 0 to K (see figure 1a). In order for the S(X) function to be concave between 0 and K, it is necessary that

$$j\mathbb{N}(X)(X-\gamma) + j(X) < 0$$

from (5). Since it is likely, especially in the education structure, that the expected penalty will be well below the social damage, γ, necessary is that j'(X) is positive, which means that the density function is increasing over the range 0 to K. This case is illustrated in figure 1b. When S(0) - S(K) < - KS'(K), it is socially desirable to tell students exactly which items will be tested. Social damage is minimized by having the students learn those and only those questions. Intuition is provided in the next section.

It is also possible that the other corner solution is optimal, where students are told that all items are subject to testing. For q=1 to be optimal, $\mathbb{M}/\mathbb{M}q$ in (8) must be negative for q=1 or

$$\frac{\partial}{\partial q} = n\left[S(K\frac{m}{n}) - S(0) - K\frac{m}{n}S'(K\frac{m}{n}) \right] < 0 \quad .$$

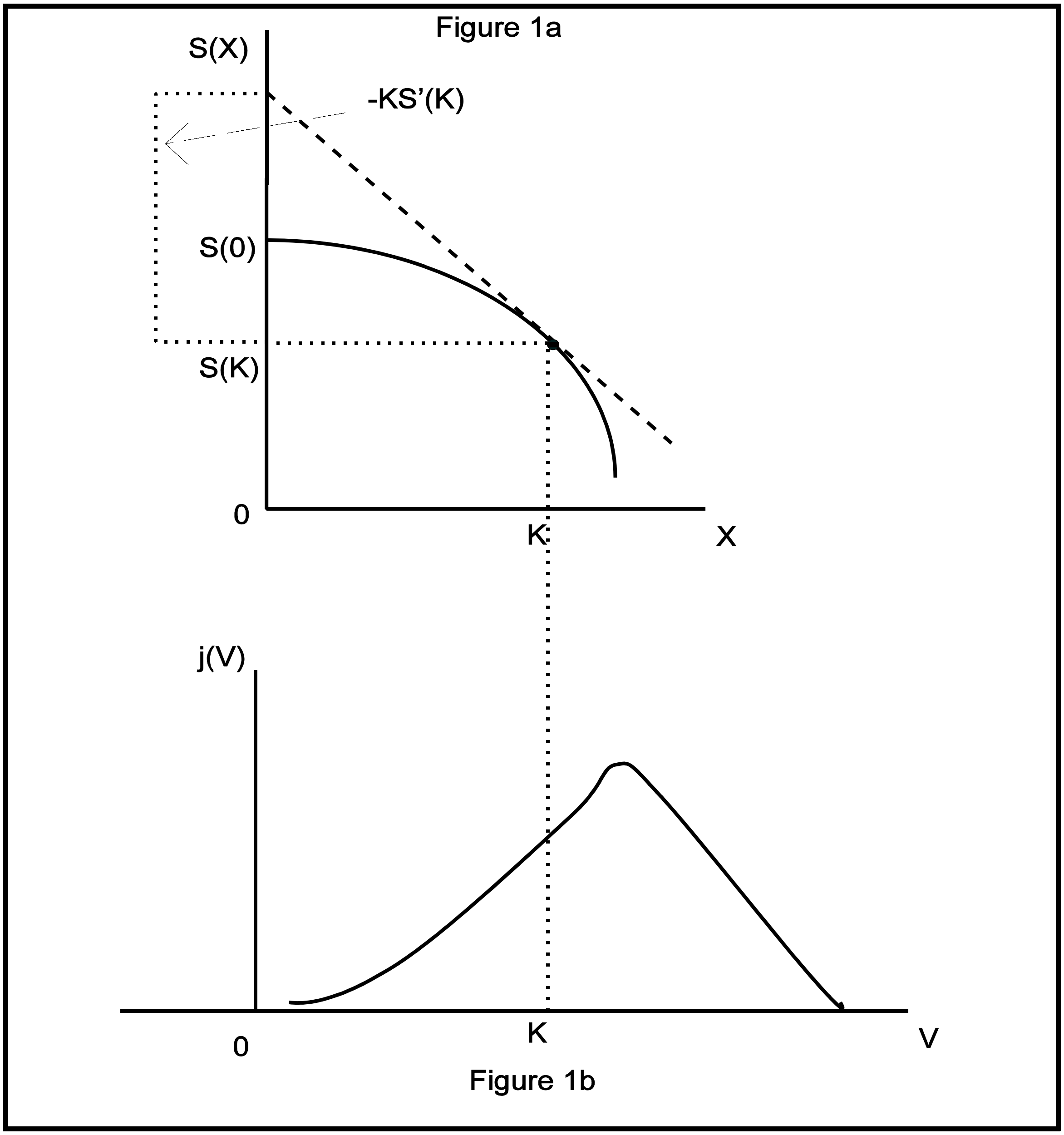


Figure 1a

Figure 1b

For this to be true, required is that

$$S(0) - S(K\frac{G}{Z}) > -K\frac{G}{Z}S'(K\frac{G}{Z})$$

which occurs if S is convex throughout the relevant range. An exponential distribution on V would be consistent with this requirement since from (5), j'<0 guarantees global convexity of S(X).

Interior solutions, where m/n<q<1, are also possible. Using (8), they occur when

$$S(K\frac{m}{qn}) - S(0) - K\frac{m}{qn}S'(K\frac{m}{qn}) = 0$$

and this can only be true for interior values of q when S(X) is neither globally concave, nor globally convex. A standard single-peaked density function on V where the peak is in the relevant range is consistent with interior solutions.

Implications and Extensions

When the social damage function is concave, it pays to concentrate the penalty on certain items in the knowledge set. When the social damage function is convex, spreading the penalty over many items minimizes the social damage. The intuition is this: Eq. (8) contains three terms. The difference between the first two,

$$S(K\frac{m}{qn}) - S(0)$$

reflects the benefit that results from having more items subject to testing. Were it not for the third term, it would be better to test more items, i.e., choose q=1. But when more items are tested, the expected penalty per item falls, resulting in less incentive per item. The net effect is ambiguous, but when the loss in incentives that results from a fall in expected penalty exceeds the gain in incentives from testing more items, it is better to have a low q.

Clearer intuition is provided by considering variations in V and in m. Variations in V reflect differences in the cost of learning. Variations in m reflect variations in the cost of testing. First consider two extreme cases on the distribution of V. In one case, let V=K for all V (across students and items). Suppose that γ>V so that it is efficient to learn all items. This can only be done if the expected penalty equals K, which means that q=m/n. The exact items to be tested must be announced. Anything short of this will result in an expected penalty less than K and therefore less than V, which means that no learning would occur.

At the other extreme, suppose that V is concentrated just below Km/n. Then, by setting q=1, that is declaring that all items are subject to testing with equal probability, students are motivated to learn because the expected penalty, Km/n, is just above the cost of learning. It would be wasteful to restrict items to any proper subset of n because then no learning would occur on the items that are not tested and nothing is gained.

The general implication is that when V is concentrated and high, it is optimal to announce the items that are to be tested. When V is concentrated and low, it is optimal to keep the questions secret.

High cost of learning requires that the questions be announced to provide sufficient incentive, whereas low cost of learning allows for secrecy because even low expected penalties induce learning.

Another intuitive result is easily derived. If it is very costly to test, it is better to announce the specific questions. It if is very cheap to monitor test, it is better to keep the questions secret.

Costly testing is reflected in a low number of questions. Suppose m is sufficiently small so that Km/n is too low a number to motivate learning, i.e., the minimum value of V exceeds Km/n. Then S(Km/n)=S(0) because failing to limit items subject to testing is equivalent to setting the penalty equal to zero; both result in no incentives. When m is sufficiently low, the optimum cannot be the policy of keeping questions secret. To see this, it is sufficient to show that the other extreme, of announcing all questions, dominates complete secrecy. It is better to announce the questions when

$$mS(K) + (n - m)S(0) < nS(K\frac{m}{n})$$

which is the same as

$$m\ S(K) + (n - m)\ S\ (0) < n\ S\ (0)$$

because X=km/n is too small to induce learning. The inequality holds because S(0)>S(K).

When questions are expensive, failure to announce them results in no learning. Announcing the questions induces learning of the announced items.

Conversely, if it is very cheap to test, then the optimum must be to leave vague the items to be tested. Suppose that m is so large that Km/n > V œ V. Formally, S(Km/n) = S(4) ; all items are learned. It is better to keep questions secret when

$$mS(K) + (n - mS(0) > nS(K\frac{m}{n})$$

which is the same as

$$\frac{m}{n}S(K) + (1 - \frac{m}{n})S(0) > S(\infty) \ .$$

This must hold because S(4) < S(X) œ X when γ>max(V).

Questions are so numerous that no student risks leaving any item unlearned. It is useful to work through a general example to show that corners, where q is equal to 1 (nothing is announced) or q is equal to m/n (the specific items to be on the test are identified), are obtained even when the V distribution is non-degenerate.

Suppose that V is distributed over the interval [0,1] with density function

j(V) = a + bV

with a and b chosen such that J(V) $ 0 and

$$\int_0^1 j(V)d(V) = 1 \ .$$

If a =0 and b=2, the density function is a triangle with mass concentrated at V close to 1. If a=2 and b=-2, the density function is a triangle with mass concentrated close to zero. If a=1 and b=0, the density is uniform.

Let n=100, m=10, γ=2 and K=0.5. Figure 2a, 2b, and 2c correspond to the situation where the density is weighted toward values of V that exceed K, i.e., where a=0 and b=2

$$S''(X) = (a-\gamma b) \quad + \quad 2\,b\,X$$

which is negative for X<1. And, as is apparent from figure 2b, the S(X) function is globally concave, which is the sufficient condition for choosing the corner where exact questions are revealed. That shows up clearly in figure 2c, where the full social damage associated with any given q (from m/n to 1) increases in q. The lowest value of q (equal to m/n) is best. It is better to reveal the questions directly so that students learn that material.

Conversely, let the density function be weighted toward low cost learning as in figure 3a. Then, as shown in figure 3b, the S(X) function is globally convex and full social damage declines in q up to 1, as is seen in figure 3c. Because there are many items that can be learned at low cost, it is better to keep the exam questions completely secret. Interior solutions have already been shown to exist in the simple case where $V=V_0$ above.

The intuition of the earlier discussion holds. When there are many items or many people who are high cost learners, it is better to announce the questions that will be on the exam. Secrecy about what is on the exam means that only very low cost items are learned. When there are many items or many people who are low cost learners, it is better to keep the questions secret. Then students will learn a larger part (although not necessarily all) of the material.[9]

Figure 2a

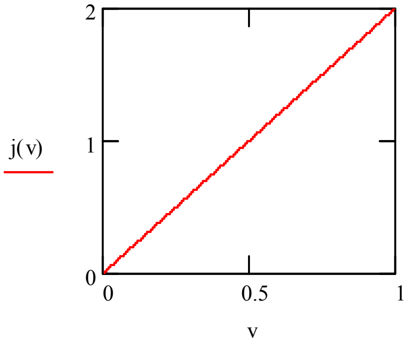


Figure 2b

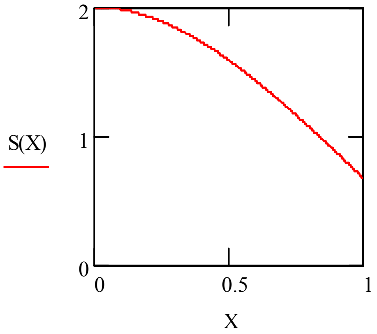
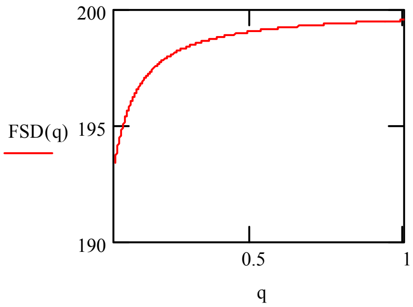


Figure 2c



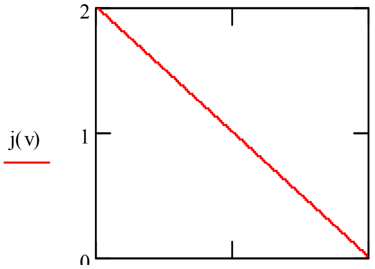


Figure 3a

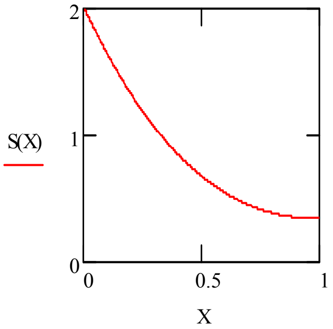


Figure 3b

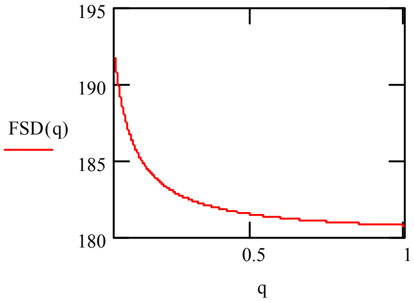


Figure 3c

What is a "Good" Test?

One common view is that 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 question selection, so as to prevent students from knowing what is on the exam. Educators often view as a goal of testing that test scores generalize to other material not on the test.[10]

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 chooses 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.

[9] The proportion of material learned is J(K), but the fraction may represent a proportion of people who learn everything, the fraction of the total knowledge base learned by each individual or a combination of the two.

[10] For example, McBee and Barnes (1998) claim that a test would have to have to test a prohibitively high number of tasks to attain acceptable levels of generalizability.

Resolving the Argument

The model captures exactly the intuition of both sides of the argument over high-stakes testing. Most agree that imposing high-stakes testing will induce teaching to the test because the incentives are strong to learn what is on the test and then to teach to it. The disagreement is over whether this is good or bad. The concern by critics of such testing is that a strategy that is tantamount to announcing the exam questions will stifle learning of the more general curriculum. These critics are correct if they have in mind students who would be sufficiently motivated to learn all the material. For high ability, low cost learners, it is possible that $V_0 < Km/n$ which implies that setting q=1 is optimal. Then, all the material is learned when all items in the knowledge base are subject to monitoring. Restricting the items that are subject to testing to a proper subset of the knowledge base wastes incentives and results in less learning than would be induced by completely open standards. But those who are the lower end of the ability distribution are in the opposite circumstance. For example, if $K=V_{0,}$ the only way to motivate any learning at all is to announce exactly which items will be tested. Setting q=m/n so that the expected penalty, (K)(m/nq), equals K results in learning of m and only m items. Because the costs of learning are so high relative to the return, students in this situation, if they are not held accountable for a smaller subset of material, 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 provides only weak 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.

Age, Background, Difficulty and Test Form

The extreme structure above is sufficient to provide intuition on why testing and monitoring methods vary across grades and schools. Consider the learning ability of young children relative to college age students. It is more costly for young children to learn academic subjects than for older ones, but probably cheaper for them to learn language.[11] As a result, V for academic subjects is higher for younger children than for older ones. The previous model says that q* should therefore be lower for younger children. At the extreme, q* = m/n so that they are told exactly what is on the test. Spelling tests given to elementary school children generally specify exactly which 10 words must be known for Friday's test. By the time students reach graduate school, only the papers and books and sometimes only the general subject area from which the test will be drawn are announced.

Analogously, children who are in honors classes are likely to have lower costs of learning

[11] There is a large body of literature on learning different skills at different ages. Perhaps best known for these ideas is Piaget and Inhelder (1969).

than those in remedial classes. Indeed, tests in honors classes are less predictable, pose questions that are extensions of material learned, and are drawn from a larger body of knowledge than those in remedial classes.

As an extension, children who attend schools in disadvantaged areas do not have the benefits of outside support that lowers the cost of learning. As a result, tests in these schools are expected to be more specific than those in high income suburbs. Under the interpretation that q=1 represents stochastic monitoring of a variety of forms, it might well be expected that the specific high-stakes tests required of disadvantaged schools would be replaced by more generic monitoring of inputs as well as outputs in high income schools.

Also, if material is known to be difficult, then for a given population of students, it is better to announce that it will be on the test. A student will only learn high V material if he knows with relative certainty that he will be tested on it. Otherwise, the expected penalty is too low to bother learning such difficult items. If the teacher expects the children to learn the toughest parts of the curriculum, she must tell them that there is a high likelihood that it will be on the exam. If she does not, the children will simply ignore the material. With easy items for which V is low, sufficient motivation to learn might be provided even if no information is given about what specifically will be on the test. As a result, it is better to be less specific about easy material.

It is also possible that learning some items makes it less costly to learn others, i.e., a learning-by-doing effect. Some pieces of knowledge, learned early, might reduce the cost of learning other items later. Then, the distribution of V becomes endogenous, depending on what has been learned before.

Now return to the more general view that V~j(V) is not concentrated at one point. Instead, V is distributed over some interval, reflecting that some material is more difficult to learn than other material, even for any given student. Also, V is not massed at one point because some students are more efficient learners than other students. As in the speeding model, the assumption is that all items in the knowledge base and all students within the same distribution given by J(V) are observationally identical. If items or people are observably different, then separate distributions must be written to characterize each, for example, as in the case of younger and older students.

As in the speeding structure, independence is assumed. Having learned one item does not affect in a direct way the cost of learning another item. This is unrealistic in three respects. First, a student may have a capacity for learning so that as the amount of studying increases, he is unable to absorb new material at the same cost. Only a limited number of items can be remembered in one sitting. Second, and working in the opposite direction is that learning begets learning. It is easier to learn calculus after algebra has been mastered. Third, because the distribution of V also includes variation across people, interactions between students is ignored. But peer pressure and identity might be important. Coleman (1961) was the first to argue this and found that the group in which an individual classified himself (e.g., bookworm, athlete...) was a predictor of academic performance. Akerlof and Kranton (2002) summarize and build on this notion to explain why academic performance may vary greatly from school to school, even when resources do not. Building in some form of dependence is possible, but complex and is ignored in this formulation.

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. 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 V 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. 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 random 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. 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, being virtually independent of performance. Intrinsic motivation is best modeled by assuming that $j(V)>0$ for $V<0$. That is, for some values of V, the cost of imparting knowledge, is negative. Even if teachers received no compensation for the amount of knowledge their students acquired, they would still choose to provide some knowledge to each student.

Under the regime of no testing, teachers would still provide J(0) knowledge to the students. The main implication is that q is likely to be larger, the more intrinsically motivated are teachers and goes to 1 for sufficiently motivated teachers. If teachers are highly motivated, then even very low expected penalties induce them to teach most if not all of the material. This is like the case illustrated in figure 3a,b,c above, where most of the V distribution is massed at the low end and well below K or even Km/n . Put differently, stochastic monitoring is relatively less effective for less motivated teachers.

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.[12]

Choosing the Right Number of Questions

Implicit in the discussion to this point was that m was exogenous.[13] 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 with announced questions. 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. Assume instead that questions can be produced at cost t(m), with t', t">0.

[12] See, for example, Holmstrom (1981) and Kandel and Lazear (1992).

[13] Hoxby (2002) outlines the economic consequences of high stakes testing. She points out that the direct costs of accountability, at least at some basic level, is very low and even the most aggressive estimates of cost do little harm to district budgets.

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

First, suppose that the optimal solution for q is interior. Then the modified full social damage function in (7) becomes

$$\text{FSD(q,m)} = n\left[qS(\frac{Km}{qn}) + (1-q)S(0) \right] + t(m)$$

The FSD(q) is modified only in that costs are recognized explicitly. The first order conditions are:

(9) a. $$\frac{\partial}{\partial m} = KS'(\frac{Km}{qn}) + t'(m) = 0$$

b. $$\frac{\partial}{\partial q} = n\left[S(\frac{Km}{qn}) - S(0) - \frac{Km}{qn}S'(\frac{Km}{qn}) \right] = 0$$

The first-order condition (9b) is as before and (9a) simply dictates setting the marginal cost equal to the social value of an additional question. Similar conditions can be derived for the corner cases, where q=m/n or q=1. The interpretations are similar.

Using (9b) and the implicit function theorem, one obtains

$$\frac{\partial q}{\partial m} = - \frac{n}{q} \frac{S'(\frac{Km}{qn})(\frac{K}{qn}) - S''(\frac{Km}{qn})(\frac{K}{qn})}{- S'(\frac{Km}{qn})(\frac{K}{qn}) + S''(\frac{Km}{qn})(\frac{K}{qn})}$$

$$= n/q \qquad > 0$$

which implies that q and m are complements in production of knowledge. This is the same result as that obtained in the speeding model. When monitoring is cheap, i.e., m is large, the optimal q rises. Exams become less well specified because more is being tested. Since testing is cheap, more knowledge can be checked and so it is better to increase the number of items subject to test.[14]

Inference and Endogenous Penalties

In determining the effect of changing the number of questions, the penalty associated with missing a question, K, has been assumed to be exogenous. That assumption is appropriate when K is indeed exogenous, as in the case when the issue is motivating teachers, not students, and K is part of the compensation policy and is constrained, say, by union or other institutional factors. It is also

[14] At the two extremes, of m=0 and m=n, all values of q provide the same incentives. When m=0, no value of q provides any incentives. When m=n, both extremes provide full incentives. The student who is told that every item in n will be tested (q=m/n=1) faces an expected penalty of K per item on every item. The student who is told that there is random sampling, but that the number of items sampled equals n faces an expected penalty of Kn/n=K again.

appropriate when under certain assumptions about production technologies. Much of this depends on the underlying statistical relationships and the derivations are tedious, but a brief discussion of the issues is provided here.

Suppose that students (rather than teachers) control their learning. Firms should "change" a penalty K for missing a question such that

$$K = E(\text{productivity} \mid (B+1)/m \text{ correct}) - E(\text{productivity} \mid B/m \text{ correct})$$

where B is some positive interger.

But inference is difficult because a correct answer could indicate skill or luck, both in innate ability and in amount learned. A general formulation would allow individuals and items to differ in cost of learning, say, as

$$V_{ij} = \delta_i + \varepsilon_{ij}$$

where δ_i is the person effect and ε_{ij} is the part of cost that is specific to the item and individual. Then the problem would be to infer δ_i as well as the amount learned given the test score. This formulation results in a different inference about changes in expected productivity associated with getting one more correct because it reflects not only the fact that a given individual knows more, but also that the individual in question is of a higher ability type.

When the agent is the teacher, rather than the student, all of this is irrelevant because K is set exogenously as part of compensation policy. Under the current system, there is little hope that information about a teacher's ability to raise students' test scores would become part of market information. If the school were free to implement an optimal compensation structure, it could easily do so. Given that the social value of learning a particular item is γ, the school would simply set $q=1$

so that all items in the knowledge set are potentially tested and choose K such that the expected penalty equals γ. If the number of questions is given as m, then the teacher would be fined K such that

$$K\, m/n = \gamma$$

or

$$K = \gamma\, n / m$$

for each question that a each of her students misses on the exam. The teacher would teach the item whenever

$$V < \text{expected fine}$$

or $$V < K\, m/n$$

or $$V < \gamma$$

which is the efficiency condition.

Monitoring Input or Output?

Formally, the model has been structured in terms of monitoring output. The monitoring may be stochastic, but it is specified in terms of output, not input. Much of the discussion of high-stakes testing views stochastic monitoring as being based on input. For example, in the absence of high-stakes tests, teachers could be monitored by having the principal visit the classroom on either a predicted or stochastic basis. As is shown here, input monitoring is accommodated by the model already presented.

Think of teachers as being in the classroom for n minutes and let one item of knowledge be conveyed if the teacher bears cost V~j(V) as before. The principal announces that he will monitor classes for m minutes (per teacher) and that q of the n minutes of total teaching time are subject to monitoring. If he finds that the teacher has not conveyed the information in the minute during which he is in the room, the teacher will be fined K for that minute in lower salary. (Of course, K may be zero or close to it.) Setting q=m/n is tantamount to telling the teacher exactly when the principal will visit the room. Setting q=1 tells the teacher that all minutes are equally likely to be monitored. Then the expected penalty is Km / qn , just as before and the teacher's decision is to teach if

$$V< Km/qn \quad .$$

Everything in the prior setup applies to monitoring on the basis of input.

Reinterpreting the model in this way means that a structure of input-based stochastic monitoring (with any level of notification of which minutes will be monitored) can be compared to high-stakes testing where all questions are announced. This simply requires comparing the expected social damage when q=m/n to that which corresponds to the current level of input-based stochastic monitoring as defined in the previous paragraph. What it does not do is *require* the interpretation that q=m/n corresponds to a new regime of high stakes monitoring and q=1 corresponds to the old system of monitoring on input. Both interpretations, of monitoring on input or monitoring on output, are consistent with any given level of q. Whether monitoring is done on the basis of input or output relates to the costs of measuring by each method and is not special to teaching. That issue has be analyzed elsewhere.[15]

[15] See, for example, Lazear (1986), and Lazear (2003).

Endogeneity of q

More generally, the issue is whether it is best to announce what is being monitored or not. A high-stakes test creates incentives for teachers and students to find out what will be tested. As such, it is closest to the case of setting q=m/n. The current alternative, which is to monitor input and sometimes output in a stochastic fashion, is formally treated at having a q>m/n and in the limit, equal to 1. Because the current situation tends to be coupled with low stakes, i.e., low values of K associated with "infractions," teachers have little incentive to attempt to discover when and how the monitoring will be done. It is for this reason that the typical situation in schools corresponds more closely to high values of q and high stakes testing to low values of q. But this argument suggests that the choice of K and of q are not independent. When the stakes are raised, there is a natural tendency by those being monitored to learn the specifics of what will be monitored, which induces a positive, endogenous relation of q to K.

Empirical Implication: Average and Variance in Test Scores

One obvious implication of the analysis is that announcing the questions raises average test scores. But the mechanism is somewhat less than obvious. The usual thought is that telling students or their teachers what will be on the exam allows them to study exactly that material, thereby raising test scores. This may be true, but it is also true that telling individuals what is on the exam induces people who would not otherwise have studied to do so. Richer implications are derived from using information on the variance in test scores.

Consider the two extreme cases already discussed. In the first case, all individuals are identical but the non-degenerate distribution of V reflects the fact that some material is more difficult to learn than other material. Suppose that test questions are unannounced. Because students are identical, they all get the same score, equal to

$$m\ J(Km/n)$$

correct. The variance in test scores is zero.

Now allow the test questions to be announced. Test scores unambiguously rise. Each student now answers m J(K) questions correctly. But again, the variance in test scores is zero.

At the other extreme, V is the same for every item in the knowledge set, but varies across individuals. When questions are unannounced, J(Km/n) of the students obtain 100% scores and [1-J(Km/n)] obtain 0. When questions are announced, J(K) obtain 100% scores and 1-J(K) obtain 0. The average rises because more students get everything right. The variance goes from

$$m\ J(Km/n)(1-J(K/m))$$

to

$$mJ(K)\ [1-J(K)] \quad .$$

Note $\frac{\partial J(1-J)}{J} = 1-2J$,

which is negative if J>0.5. Variance falls as average test scores rise if the proportion who get all right is greater than 50%.

The true situation is neither extreme, but the implication by continuity is that announcing the questions raises average test scores because any given individual’s incentives to learn the designated items rise and because more individuals opt to study in the first place. The variance in test scores may go up or down, depending on the relevant proportions.

Additionally, it would be possible to estimate the underlying distributions, J(V) (given some sufficiently concrete parameterization) by seeing how the mean and variance of test scores change when the amount of information about the questions to be on the exam is changed.

Test Design and Learning Incentives

It is possible to ask how test design, and in particular scoring, affects incentives. For example, one very large, high-stakes test could be given or many smaller, low stakes tests could be required. The incentives to study and/or teach are very different under the two approaches.

Additionally, exams could be graded pass/fail or in a continuous fashion. The pass/fail structure is more like a tournament against a standard, where the standard is calibrated on the basis of previous classes' performances. A continuous grading structure is like paying a piece rate. It is already known that the incentive effects of the two different structures vary, depending on the nature of the payoff scheme and the heterogeneity of the underlying population.[16]

On a different note, good exams are neither too easy nor too difficult, and this is primarily for statistical reasons but also because of the effect on incentives. On a very easy exam, a careless mistake can cause a student to fall well below the rest of the class. Such exams have low signal-to-noise ratios. On a very difficult exam, average scores are very low, and it is difficult to distinguish among people because all do so poorly. Again the signal-to-noise ratio is low. As for incentives, it

[16] See Lazear(2000) and Lazear and Rosen (1981).

is a general principle in incentive theory that when noise is high relative to the signal, incentives are diminished. The optimal test difficulty should take this incentive effect as well as statistical issues into account.

Investigation of these issues is left to subsequent work.

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.[17] 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 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 of Harvard makes the point

[17] See 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).

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. 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.[18] 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 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

[18] Again, see Koretz, et. al. (1991).

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.

For these reasons, this analysis has assumed that each of the n items in the knowledge base are perfect substitutes for one another in testing. Although not literally true, this is likely to be a good approximation for the issue that is central to the policy debate. Most important, unlike earlier analyses, the approach here provides insights into how to structure the socially efficient incentive scheme.

A Theoretical Observation on Optimal Enforcement

The model used above is a generalization of the usual optimal enforcement literature. Rather than having the probability of detection being the same over all actions, we allow for different probabilities of detection. In the case of education, it is that qn items have a probability of detection equal to m/qn and (1-q)n items have a probability of detection of zero. This generalization can be extended. At the extreme, it is possible to allow for each item (or each mile) to be given a different probability of detection. Enforcement resources could be spread over the entire set of possible crimes so as to minimize social loss.[19]

Different characteristics lead to different strategies. It has already been shown that when detection is expensive or costs of learning or teaching are high, it is best to go to a corner, having the questions announced in advanced. Since it is only the difference between social value and social

[19] I thank Thomas Dohmen for pointing out this generalization.

cost, γ-V, that is relevant it is possible to turn this around and think of situations where the difference is high because γ is high, not because V is low. For example, it might be optimal to put the police on more congested roads where the social cost of an accident is especially high. Similarly, it might be better to test certain skills like basic literacy over others because they have higher social value. The model can be interpreted to address these issues, but more structure would be needed before specific implications about items tested could be provided.

Conclusion

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

The analysis has shown that providing well-defined requirements dominates stochastic incentives for individuals for whom compliance costs are high. In the context of education, this means that predictable 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.

Put differently, 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 for failing teachers, tests that are predictable are best at providing incentives to learn. For high ability students or successful teachers, somewhat less predictable tests are best.

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. Number of questions and randomness are complements. When testing is cheap, the optimal number of questions rises. At the same time, the proportion of items which are subject to testing rises. Put differently, if it is very cheap, then it is better to keep the nature of the exam highly secret.

3. Exam specifics are made known to younger children and for difficult material because revealing exam questions provides better incentives.

References

Akerlof, George A. and Rachel E. Kranton. "Identity and Schooling: Some Lessons for the Economics of Education," *J. Econ. Lit.*, XL, 1167-2001, December, 2002.

Baker, George. "Incentive Contracts and Performance Measurement." *Journal of Political Economy* 100:3 (June 1992): 598-614.

Becker, Gary. "Crime and Punishment: An Economic Approach." *Journal of Political Economy* 76 (March-April 1968): 169-217.

Coleman, James. *The Adolescent Society: The social Life of the Teenager and Its Impact on Education.* NY: The Free Press, 1961.

Heckman, James, Carolyn Heinrich, and Jeffrey Smith. "The Performance of Performance Standards." *Journal of Human Resources* 37:4 (Fall 2002): 778-811.

Hoffman, James V., Lori Czop Assaf, and Scott G. Paris. "High-stakes testing in reading: Today in Texas, tomorrow?" *The Reading Teacher* 54:5 (Feb 2001): 482-492.

Hölmstrom, Bengt. "Contractual Models of the Labor Market." *American Economic Review Papers and Proceedings* 71 (1981): 308-13.

Holmstrom, Bengt and Paul Milgrom (1991). "Multi-task Principal-Agent Analyses: Incentive Contracts, Asset Ownership and Job Design." *Journal of Law, Economics, and Organization* 7 (1991): 24-52.

Hoxby, Caroline M., "The Cost of Accountability," in Williams Evers and Herbert Walbert, eds., School Accountability. Stanford Hoover Press, 2002.

Jones, M. Gail, Brett D. Jones, Belinda Hardin, Lisa Chapman, Tracie Yarbrough, and Marcia\ Davis. "The Impact of High-Stakes Testing on Teachers and Students in North Carolina." *Phi Delta Kappan* 81: 3 (November 1999): 199-203.

Kandel, Eugene and Edward P Lazear. "Peer Pressure and Partnerships." *Journal of Political Economy* 100:4 (August 1992): 801-817.

Koretz, Daniel M., Robert L. Linn, Stephen B. Dunbar, and Lorrie A. Shepard. "The Effects of High-Stakes Testing on Achievement: Preliminary Findings about Generalization Across Tests." Originally presented in R. L. Linn (Chair), *Effects of High-Stakes Testing on Instruction and Achievement*, symposium presented at the annual meeting of the American Educational Research Association and the National Council on Measurement in Education, Chicago, April 5, 1991.

Lazear, Edward P. "Salaries and Piece Rates." *Journal of Business* 59 (July 1986): 405-31.

Lazear, Edward P. "The Power of Incentives." *American Economic Review* 90:2 (May 2000): 410-414.

Lazear, Edward P. "Teacher Incentives." *Swedish Economic Policy Review* 10:2 (2003): 179-214.

Lazear, Edward P. and Sherwin Rosen. "Rank-Order Tournaments as Optimum Labor Contracts." *Journal of Political Economy* 89:5 (October 1981): 841-64.

McBee, Maridyth M., and Laura L. B. Barnes. "The Generalizability of a Performance Assessment Measuring Achievement in Eighth-Grade Mathematics." *Applied Measurement in Education* 11:2 (1998): 179-194.

Meisels, Samuel J. "High-Stakes Testing in Kindergarten." *Educational Leadership* 46:7 (April 1989): 16-22.

Piaget, Jean and Bärbel Inhelder. *The Psychology of the Child.* Trans. by Helen Weaver. New York: Basic Books, 1969.

Rivkin, Steven G., Eric A. Hanushek, and John F. Kain. "Teachers, Schools, and Academic Achievement." National Bureau of Economic Research Working Paper No. 6691 (revised 2001).

Appendix

In what follows, speed, σ, is a choice variable. This means that the benefit from speeding, V, must be made a function of speed. Thus, let

$$V = V(\sigma, \theta)$$

where θ us a random variable with density a(θ) that varies across people and miles so that the value of speeding depends not only on the speed, but also on the person and on the specific mile.

Similarly, the social damage depends on speed so γ is replaced by

$$\gamma = \gamma(\sigma)$$

Finally, the probability of being caught depends on the speed chosen so

prob of being caught = p(σ)

On any given mile, a driver, who knows his θ, chooses σ to maximize expected utility given by

(A1) Expected Utility = V(σ,θ) - p(σ) (KG) / (qZ) on patrolled miles

and

= V(σ,θ) on unpatrolled miles

The first order conditions are

(A2) $$\frac{\partial}{\partial\theta} = V_1 - p'(\sigma)(KG)/(qZ) = 0$$ on patrolled miles and

$$\frac{\partial}{\partial \theta} = V_1 = 0 \quad \text{on unpatrolled miles}$$

(On unpatrolled miles, the driver limits his speed even if there were zero probability of a fine simply because of the danger and disutility associated with speed beyond some point.) Define the solutions to (A2) to be σ*(q,θ) and $\sigma_0(\theta)$, respectively.

Then, the full social damage function is given by

$$FSD(q) = Z\Big\{ q\int [\gamma(\sigma^*(\theta,q)) - V(\sigma^*(\theta,q),\theta)]a(\theta)d\theta + (1-q)\int [\gamma(\sigma_0(\theta)) - V(\sigma_0(\theta),\theta)]a(\theta)d\theta \Big\} \tag{A3}$$

As before, to reveal exact locations, it is necessary that FSD'(G/Z) > 0 and a sufficient condition is that. In order to provide no information about where police are located, it is necessary that FSD'(1)<0 and a sufficient condition is that FSD'(q)<0 for G/Z < q $\leq$ 1.

It is possible to differentiate (A3) to try to determine under which conditions corners are generated. But there are so many free functions in (A3) that little can be said that is of much value.