Speeding, Tax Fraud, and Teaching to the Test

Edward P. Lazear

Hoover Institution
and
Graduate School of Business

Stanford University

June, 2004

This research was supported by CRESST. I am indebted to Edward Glaeser for some of the derivations contained in the paper. In addition, Gary Becker, 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, 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 teaching monitoring is 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, but speed on 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 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 likely deter cheating on those items, which is better than failing to deter any cheating.

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

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 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

$$K G / Z .$$

[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.

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

(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. As a result, if the location of the police 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 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, so it is 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 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.

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. 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) $K G / Z < V < K$,

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. 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.

Optimal Deterrence

Assuming that it is always optimal to deter speeding, but not to deter tax fraud makes clear that it is important to specify the social costs and benefits of each action. When the focus later turns to education, this will become even more central. Therefore, we return to the speeding example, but model social costs explicitly and drop the assumption that V is identical for all people for all miles.

Allow there to be a distribution of V that reflects the value of speeding on any given mile by any given person. Let that distribution be written J(V) with corresponding density j(V). The unit of analysis is a person mile so that V can vary for a given individual because speeding on some miles or at some times is more valuable than others and V can vary across people so that some people place a higher value on speeding than others. Then j(V) is the density across all miles driven by all drivers. Note that this structure means that a given driver might speed on some roads and not on others and some drivers might speed sometimes or always and others never, depending on the distribution of V across miles and people. Further note that the assumption is that all miles and drivers characterized by the distribution J(V) are observationally identical. If miles or people are observably different, then separate distributions must be written to characterize each. Finally, note that V are assumed to be independent of whether speeding occurs on other miles. For the sake of simplicity, such complications are ignored.

Suppose that there is some expected penalty, X. A given driver speeds on a given mile if and only if V>X. Let the social cost of speeding by given by γ. Individuals for whom V<X do not speed. Those for whom V>X speed. They receive V in benefit and impose cost γ on society. Thus, the social damage associated with any given fine 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) $S''(X) = j'(X)(X-\gamma) + j(X)$

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 $X=\gamma$. Setting the expected fine equal to the social cost of the infraction induces the appropriate behavior.

Continuous Choice and Interior Solutions

In the simple model, the choice was between two alternatives. Either drivers were told that police were stationed randomly over all existing roads or they were told that 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. Specifically, given that there are G police, drivers can be told that there is some proportion of all roads, q, that are patrolled and some proportion, 1-q, that are not patrolled. The California Highway Patrol uses exactly this strategy. For example on the July 4 weekend of 2004, they 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 for patrols to check for intoxicated drivers.

On the patrolled roads, the probability that any one mile is patrolled is not necessarily 1. In general, since qZ miles are subject to patrol and G are actually patrolled, the probability that a mile is patrolled, given that it is in the set of potentially patrolled roads is G / qZ. The case where police are distributed randomly over all the roads is merely a special case given by q=1, where the probability of any one mile being patrolled is simply G / Z, as before. The other extreme case, where drivers are told exactly which miles are patrolled is given by q=G/Z. Then, on patrolled roads, the probability of any given mile being patrolled is 1 and it is zero on all other roads.

The problem then is to choose q so as to minimize social damage. It is shown in this section that extreme cases are possible where the solutions are q= G/Z (tell where the police are) or q=1 (do not tell where the police are).

Expected damage as a function of q is given by

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

On the q Z patrolled miles, damage is S(X) and on the (1-q) Z unpatrolled miles, the damage is S(0). The expected penalty on the patrolled miles is

$$X = K\frac{G}{q\,Z}$$

so (6) can be written as

(7) Full social damage = Z [q S($K\frac{G}{q\,Z}$) + (1-q) S(0)]

Differentiate with respect to q to obtain

$$(8) \qquad \frac{\partial}{\partial q} = Z\left[S(K\frac{G}{qZ}) - S(0) - K\frac{G}{qZ}S'(K\frac{G}{qZ}) \right]$$

In order for it to be optimal to state exactly where the police are located, q must be equal to G/Z. This happens if $\partial/\partial q$ in (8) is positive at q=G/Z. Then, increasing q will only increase social damage so the corner solution is best.[5] The requirement is that

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

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

$$j'(X)(X-\gamma) + j(X) < 0$$

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. Under these circumstances, it is socially desirable to tell drivers exactly on which miles police are located. Social damage is minimized by having the speed law obeyed on those miles and nowhere else.[6]

It is also possible that the other corner solution is optimal, where drivers are told that police are equally likely to be found on each of the Z miles, which implies q=1. For q=1 to be optimal,

---

[5]An additional requirement is the regularity condition that if $\partial/\partial q$ becomes negative, S(K)>S(KG / qZ).

[6]It is still true that some (those for whom V>K) will choose to speed.

M/Mq in (8) must be negative for q=1 or

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

Figure 1a

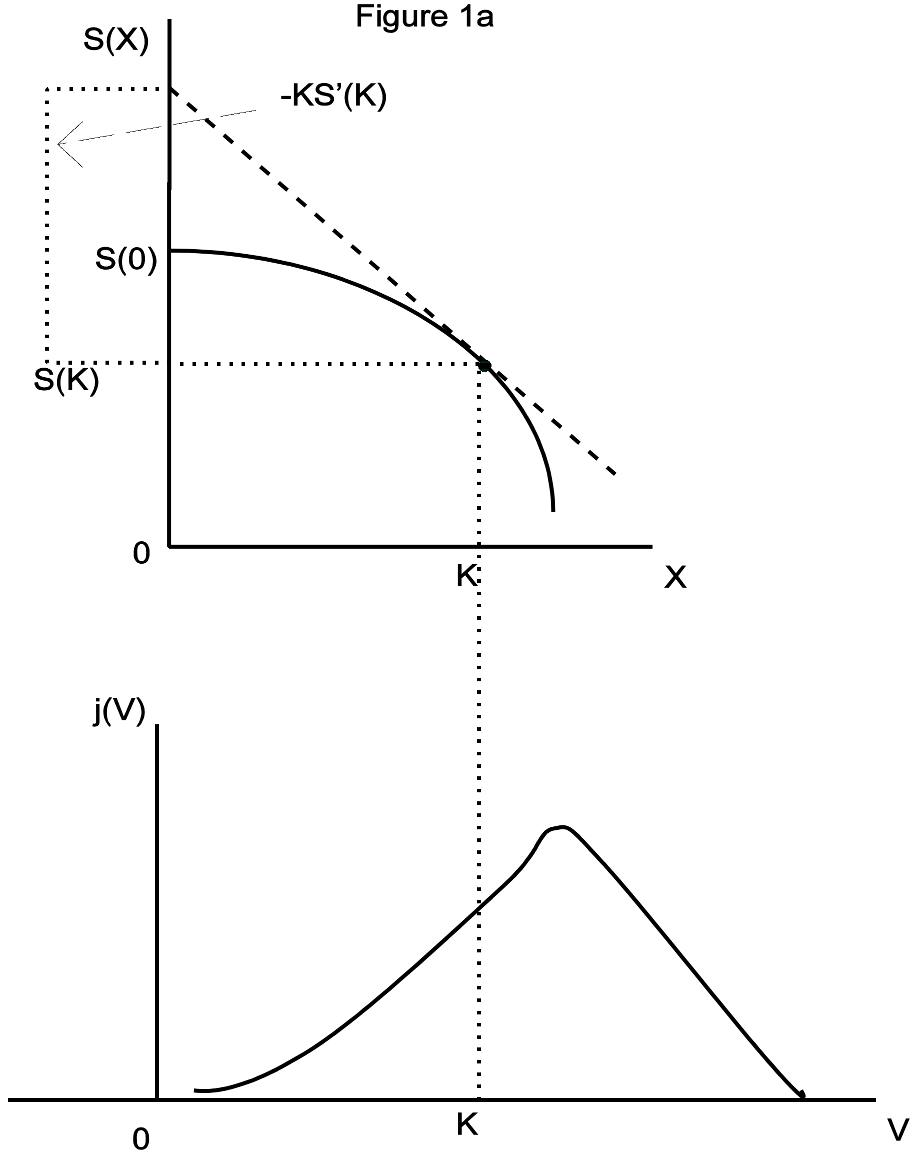


Figure 1b

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 G/Z<q<1, are also possible. Using (8), they occur when

$$S(K\frac{G}{qZ}) - S(0) - K\frac{G}{qZ}S'(K\frac{G}{qZ}) = 0$$

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 a few miles of road. When the social damage function is convex, spreading the penalty over many miles of road minimizes social damage. The intuition is this: Eq. (8) contains three terms. The first two,

$$S(K\frac{G}{qZ}) - S(0)\ ,$$

is the gain that results from having more miles subject to patrol and, therefore, fine. It is negative when it is efficient to deter speeding because a positive fine deters more speeding than a zero fine, resulting in lower social damage. Were it not for the third term, higher q, meaning more miles patrolled would be better. But when more miles are patrolled, the expected penalty per mile falls, resulting in less deterrence per patrolled mile. The net effect is ambiguous, but when the loss in deterrence that results from a fall in expected penalty exceeds the gain in deterrence from patrolling more miles, it is better to have a low q.

Clearer intuition is provided by considering variations in V and in G. Variations in V reflect differences in the value of speeding. Variations in G reflect variations in the cost of detection. First consider two extreme cases on the distribution of V. In one case, let V=K for all V (across people and miles). Suppose that γ>V so that it is efficient to deter speeding everywhere and for everyone. This can only be done if the expected penalty equals K, which means that q=G/Z. The exact miles on which police are located must be announced. Anything short of this will result in an expected penalty less than K and therefore less than V, which means that speeding would occur on all roads. Identifying exactly all patrolled miles results in the law being obeyed on G miles.

At the other extreme, suppose that V is concentrated at epsilon below V=KG/Z. Then, by setting q=1, that is declaring that all miles are subject to patrol with equal probability, drivers are just deterred from speeding on all miles because the expected penalty, KG/Z, is just above the value of speeding. It would be wasteful to restrict patrolled miles to any proper subset of Z because then speeding would occur on the unpatrolled miles and there is no reason to permit that to occur.

The general implication is that when V is concentrated and high, it is optimal to announce the location of the police. When V is concentrated and low, it is optimal to keep the locations secret.

High values of speeding require that the locations be announced to provide deterrence, whereas low values of speeding allow for secrecy because even low expected penalties deter speeding.

Another intuitive result is easily derived. If it is very costly to monitor speeding, it is better to announce the location of the police. It if is very cheap to monitor speeding, it is better to keep the location secret.

Costly monitoring is reflected in a low number of police. Suppose G is sufficiently small so that K G/Z is too low a number to deter any speeding, i.e., the minimum value of V exceeds KG/Z. Then S(GK/Z)=S(0) because failing to limit the area patrolled is equivalent to setting the fine equal to zero; both result in no deterrence. When G is sufficiently low, the optimum cannot be the policy of keeping locations secret. To see this, it is sufficient to show that the other extreme, of announcing exact locations, dominates complete secrecy. If the choice is one of the two extremes, i.e., announcing the exact location of the police and not providing any information on their location, then it is better to announce the exact location when

$$G\,S(K) + (Z-G)\,S(0) < Z\,S(K\,G/Z)$$

which is the same as

$$G\,S(K) + (Z-G)\,S(0) < Z\,S(0) \quad .$$

because G is so small that there is no deterrence at all when locations are secret so S(K G/Z)=S(0). The inequality holds because S(0)>S(K).

The idea is that because police are so few, failure to announce their locations results in speeding everywhere. At least if their locations are known, speeding will be deterred on the patrolled sections.

Conversely, if it is very cheap to provide police and it is optimal to deter all speeding (γ>max(V)), then the optimum must be to leave the location of the police completely secret. If it is cheap to monitor, then G is very large. Suppose that G is so large that K G/Z > V for all V. That is, because there are so many police, there are no miles on which speeding occurs even when location of patrolled miles is a complete mystery. Formally, S(K G/Z) = S(4) ; all speeding is deterred. It is better to leave the location of police secret when

$$G\,S(K) + (Z-G)\,S(0) > Z\,S(K\,G/Z)$$

which is the same as

$$\frac{G}{Z}S(K) - (1-\frac{G}{Z})S(0) > S(\infty)\ .$$

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

The idea is that police are so abundant that no driver will risk speeding on any mile. The likelihood of being caught is sufficiently great to deter speeding on all roads, given the fine.

Up to this point, the driver is given two choices only: speed or obey the law. It is possible, to allow the speed chosen to be continuous and to allow the probability of the fine, the damage associated with speeding, and the utility from speeding to depend on speed in a smooth way. In the context of teaching to the test, rather than assuming that the material is learned or not, the continuous analogue would permit different levels of learning of a given item of knowledge. The cost and benefit, both social and private, would vary with the amount of learning that occurs. The conclusions and implications are no different from the ones already derived. The functions are more complex, but the analysis is the same. In order to announce the miles patrolled with certainty, it is necessary that FSD(q) is increasing in q when q=G/Z. In order to give no information about which miles are patrolled, it is necessary that FSD(q) is decreasing in q when q=1.

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 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 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. Second, high-stakes testing with well-defined exam questions is best used when the distribution is weighted toward high cost learners.[7]

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, testing as a policy issue 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. 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

[7]The emphasis here is on the incentive aspect of testing. Another role of testing is to inform teachers of what students do not know so that the curriculum can be modified. To deal with this component of testing, a dynamic model is required.

in more detail. In this section, exogeneity of K is assumed for convenience. But when the problem is shifted to motivating teachers, K is not market determined and is best thought to be exogenous.

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 $0\leq m\leq n$. There are two interpretations. The first and most direct is that the student is tested on m items, but they are drawn randomly from the body of knowledge n. The second interpretation is that there might not be any high-stakes test at all, but the student is still monitored with an intensity equivalent to testing m/n items. To make the comparison appropriate, it is necessary the similar intensities of monitoring across the two regimes are compared.[8] Such monitoring could be on input or output. Students could be randomly questioned about material that they had learned. Alternatively, teaching methods could be monitored on a random basis, as could

[8]The 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.

teacher knowledge and curriculum. All of these are meant to be proxied by the stochastic monitoring intensity given my m/n. If m=n, then the monitoring intensity is such that all items in the knowledge set are monitored.[9]

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 at least learn the m items when there is an announced, high stakes test.

[9]One 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.

It is possible to do better than both extremes, which are in fact special cases of the general formulation. As before, let us announce that q of the n items in the knowledge base are subject to testing. In this simplest case, where $V=V_0$ across all items and all individuals, the minimum expected penalty that will induce an individual to learn is one such that the expected penalty exactly equals V_0. Then, to solve for the optimal q it is only necessary to find q such that

(9) $K\,m/(nq) = V_0$

or equivalently,

(10) $q^* = (K/V_0)(m/n)$

Note that q* is the maximum value of q that induces learning. High values of q mean that the expected penalty for not knowing any given item is reduced. Setting q=q* allows the largest number of items in the knowledge base to be learned because incentives are just sufficient to induce learning on qN items when the expected penalty on these items is equal to V_0. The expected penalty on all other items, which are announced to be off the test, is zero.

Put in terms of the social damage structure of the speeding model,

$S(X) = \gamma$ for $X<V_0$.

$S(X) = V_0$ for $X \geq V_0$.

The full social damage as a function of q is then

$$FSD(q) = q\,\gamma + (1-q)\,\gamma \quad \text{for } q > q^*$$

$$= \gamma \quad \text{for } q > q^*$$

and

$$FSD(q) = q\,V_0 + (1-q)\,\gamma \quad \text{for } q \leq q^*$$

It is better to set $q \leq q^*$ because $V_0 < \gamma$. Differentiating FSD(q) with respect to q on the relevant branch gives

$$M/Mq = V_0 - \gamma$$

which is negative. Thus, full social damage is minimized by going to the point where $q=q^*$.

In general, the expected fine that the student pays in lost wages is not equal to the social cost for two reasons. First, there may be positive externalities (reduced crime, welfare dependence, and other social problems) associated with education. Second, the market's penalty in the form of lower wages depends on the statistical model. For example, if students were identical and differences in exam scores merely reflected randomness associated with question selection, there would be no reason for the market to penalize any student who missed an exam question. But then students would pay no price for missing questions and incentives would be reduced to zero.

The model produces some implications. Using (10), note that

$$\frac{\partial q}{\partial V_0} = \frac{-Km}{nV_0^2}$$

which is negative. The optimal q decreases as V_0 increases. As learning becomes more costly, it is necessary to be more precise about the material for which the student will be held responsible. High cost learners will simply give up and learn nothing if they are told that the amount of material over which they will be tested is too large.

Also from (10),

$$Mq^*/Mm = K/(nV_0)$$

and

$$Mq^*/Mn = -(K / V_0)(m/n^2) .$$

The cheaper is testing, reflecting in a higher value of m, the larger is the optimal q*. As monitoring costs fall, it is possible to provide appropriate learning incentives on a large number of items in the knowledge base. Conversely, as the number of items in the knowledge base increases, it is more difficult to provide incentives because there are more pieces of information from which to draw. To keep incentives sufficient, it is necessary to reduce the proportion of the total knowledge base that is tested so that the absolute number of items subject to test (i.e., q n) remains unchanged.

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, avoiding the type of problems illustrated by the example of testing regular polygons but never testing irregular polygons. Educators often view as a goal of testing that the 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.

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.

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

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 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.

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 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

[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).

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 more vague 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. The form of testing and resulting incentives would also depend on what was previously learned.

A General Formulation

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 two 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 at 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. Building in some form of dependence is possible, but complex and is ignored in this formulation.

Given that q of the n items are subject to test, the amount learned is

$$q\,n \int_0^{\frac{Km}{qn}} v\, j(v)\, dv$$

$$= q\,n\,J(\frac{Km}{qn})$$

But the social problem is not to maximize the amount learned because learning carries with it a social cost of V and a social value of γ. Instead, the problem is one of maximizing social surplus or equivalently, minimizing full social damage.

A re-writing of (7) and (8) above to accommodate the notation used for the teaching context gives

(11) Full social damage = n[q S($K\frac{m}{q\,m}$) + (1-q) S(0)]

and

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

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 Km/n to 1) increases in q. The lowest value of q (equal to Km/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 section 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

Figure 2a

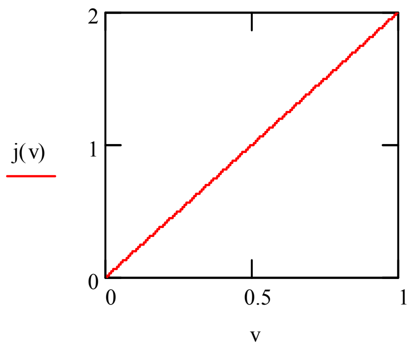


Figure 2b

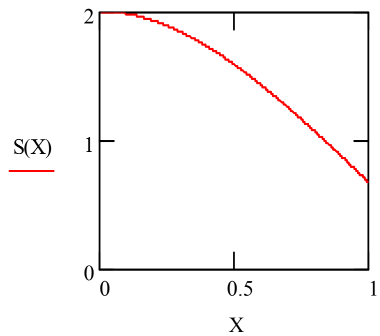
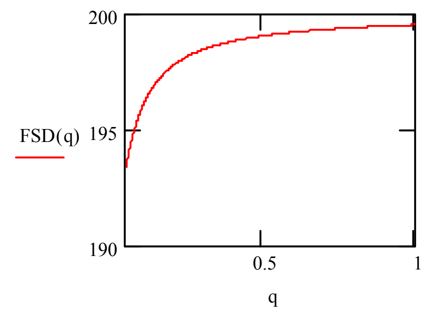


Figure 2c



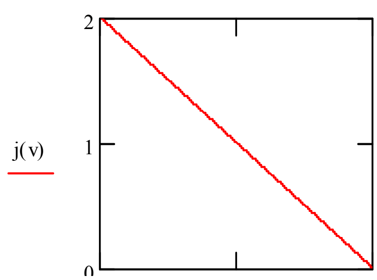


Figure 3a

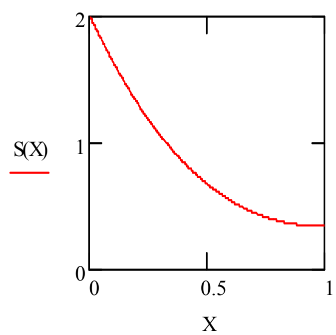


Figure 3b

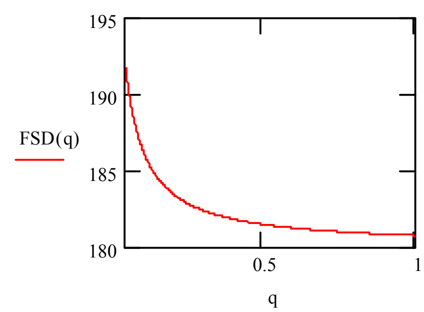


Figure 3c

learn a larger part (although not necessarily all) of the material.[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.

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 (11) 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:

[12]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.

[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.

(13) 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 (13b) is as before and (13a) 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 (13b) and the implicit function theorem, one obtains

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

$$= n/q > 0$$

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

Suppose that students (rather than teachers) control their learning. Then K should reflect an inference that employers draw about their ability on the job and how that inference is affected by missing a question. At one extreme, assume that all individuals are alike, but that V is non-degenerate because knowledge items are of differing difficulty. Each student makes an identical study to learn only and all of those items for which $V < Km/qn$. Equilibrium requires that K be set such that K equals the value of knowing one more item. Let h denote the number of items that each student learns.

---

[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.

Then employers will be willing to pay an additional K for each additional item learned such that

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

Given that all students are alike, every student makes the same choice, which is to learn up items up to the point where $V = Km / qn$. Thus, the number of items learned is

$$h^* = n\, J(Km / qn)$$

because $J(Km/qn)$ is the proportion of the n questions such that $V < Km/qn$.

A simple example makes the point. Let productivity be given by

$$\text{Productivity} = h/2.$$

Then K= ½. Given K= ½, the student chooses

$$h^* = n\, J(m / 2qn)$$

Using the example in fig 3a,b,c, where m=10, n=100, and the distribution of v is $j(V) = 2-2V$,

$$h^* = 9.75.$$

Were m to become cheaper such that now eleven instead of ten questions were on the exam, the equilibrium number of items learned would go up by slightly less than one unit, to 10.725. But K would remain constant at ½ because the effect of getting one more correct (which does not imply knowing one more item) would still be given by ½ because of the assumed relation of productivity to knowledge.

In general, the situation is more complex. A more general formulation would allow both 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 δi 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.

The main conclusion to draw from this section, however, is that the assumption of a constant K is consistent with one structure in a competitive labor market setting and may be taken as a first approximation to reality.

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. If the social value of learning a particular item were γ, then 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

$V <$ expected fine

or $V < K\, m/n$

or $V < \gamma$

which is the efficiency condition.

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 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. This is true even when the size of the penalty and monitoring intensity is the same in both regimes.

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.[15]

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

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.[16]

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 current situation corresponds more closely to high values of q and high stakes testing to low values of q. If true, then 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.

---

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

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 of 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.[17]

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. It 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 fairness 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.[18]

---

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

[18]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

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 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.[19] 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.

---

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).

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

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 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.

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 more unpredictable 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. If testing is very cheap, then it is better to keep the nature of the exam highly secret.

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

References

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.

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 $a(\theta)$ 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 = $p(\sigma)$

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

(A1) Expected Utility = $V(\sigma,\theta) - p(\sigma)(KG)/(qZ)$ on patrolled miles

and

$= V(\sigma,\theta)$ on unpatrolled miles

The first order conditions are

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

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

(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 $\sigma^*(q,\theta)$ and $\sigma_0(\theta)$, respectively.

Then, the full social damage function is given by

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

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.