Theory and Research in Education, 2, 2004, 271-281.

HIGH STAKES

Catherine Z. Elgin

Harvard University

Abstract: I discuss the contributions of Harvey Siegel, Francis Schrag, and Randall Curren to this volume. Their papers cast in bold relief the relation of High Stakes Testing to the goals of education, the nature of mind, and the demands of justice. I argue that the connections are deep but that the considerations these authors raise do not show that High Stakes Tests are in principle unacceptable. Rather they show that we need to be exceedingly careful about how our assessments are constructed, how the results are interpreted, what we take them to reveal, and what we do with the results.

Keywords: high stakes testing, assessment, justice, other minds.

Promotion and graduation in the United States increasingly depend on the results of High Stakes Tests. So do property values and college prospects. Indeed, our leaders tell us, the political and economic fate of the nation hangs in the balance. All of this is familiar. The papers in this volume show that more is at stake than this: the role of High Stakes Testing in education brings into relief the goals of education, the nature of the mind and the demands of justice. The questions that concern me are 'in principle' questions. They do not focus any particular test, but ask whether there could be a test of a given kind that does what we want such a test to do. To answer this requires getting clear about what kind of test we are talking about and what we want such a test to do.

A number of issues tend to get conflated in discussions of High Stakes Testing. (1) Should there be High Stakes Tests? That is, is it appropriate that significant academic consequences turn on one's performance on a single test? This is a question about the stakes. (2) If such tests are ever appropriate, what should the content of the tests be? That is, what sorts of abilities or accomplishments should or can students be expected to manifest on a High Stakes Test? (3) What testing format is capable of disclosing whether students have the requisite abilities or accomplishments? (4) How much of what is educationally valuable can be tested by such tests?

The three papers focus on different aspects of the problem of High Stakes Testing.[1] Harvey Siegel contends that the public justification (at least in Florida) is economic, and argues that such a justification is both skewed and narrow. The idea that the overriding goal of education is to prepare students for the workforce does each of them an injustice. Francis Schrag focuses on issues having to do with distributive justice. He accepts, at least for the purposes of argument, an economic focus, and considers how various distributions of test results might foster distributive justice by affecting the proportions of black and white high school students who graduate. Randall Curren construes the problem of High Stakes Tests as problem of other minds. He is concerned with whether standardized tests afford the sort of access to student minds that enable us to tell whether they understand the material. His question directly concerns the type of test, not the stakes. But if, as some contend, standardized tests are necessarily superficial, it seems obvious that we should not place much weight on their results.

Florida's public justification for High Stakes Testing is economic. Students who pass the Florida Comprehensive Achievement Tests (FCATs), it is said, will at a

minimum be less likely to end up on welfare, and on a more optimistic view, contribute to the state's economic prosperity. Siegel objects that this betrays a constricted and distorted conception of education. It takes the role of education to be instrumental, and justifies the tests on the grounds that they conduce to some external (and not wildly attractive) end. As Schrag points out, the economic advantages of graduating from high school are considerable. So we should not discount the economic argument completely. But to represent the entire, or the main, justification as economic is clearly wrong. Education contributes to and figures in multiple values, both intrinsic and extrinsic. It is impossible, I believe, to answer the question 'What is a good education?' without at least implicitly answering the question 'What kind of life is a good life, all things considered?'

Still, before giving up on Florida, a couple of questions need to be asked. One is what passing the FCAT is supposed to show. If it is supposed to demonstrate that a student is an adequately educated high school student, and the test is in fact designed merely to insure that those who pass it are employable, Siegel is clearly right. But if it is supposed to indicate only one aspect of what makes for an adequately educated high school student (call it the 'at-least-he-can-hold-down-a-job' aspect), the test may be reasonable. Other measures would be used to assess other factors.

Another question is whether the public justification is the real justification. Do the legislators and the state department of education think that economic considerations provide the only or the main justification for the tests and the education they bear on? Or do they merely believe that the economic argument is an argument that the voters can understand and get behind? As Siegel notes, there are multiple, plausible educational ideals: fostering creativity or critical thinking or autonomy, maximizing freedom or

individual happiness or group solidarity, reinforcing religious faith or democratic ideals, and so forth. Many of these ideals are highly contested. Some parents do not endorse them for their own children; some taxpayers are unenthusiastic about paying for other people's children to realize them. So justifying public education in terms of these ideals is politically risky. But pretty much everyone endorses the idea that graduates of the public schools should have the capacity to support themselves. The economic conception might function as a least common denominator in the political calculations of the state department of education.

I do not know anything about the FCAT or what went into its design, but I did pay some attention to the debates about what the Massachusetts Comprehensive Achievement System (MCAS) should cover. There were substantive, sometimes quite bitter debates about what students should know about history, science, math, and literature. The arguments for (and against) insisting on knowledge of, for example, *Romeo and Juliet,* the Council of Trent, the life cycle of a star, and so forth were not economic. The disagreement about whether *Silas Marner* should be on the list was not about whether that reading the novel would have a desirable effect on economic productivity. All parties to the debates had rich, multi-faceted conceptions of the ideals of education. What made the disputes so hard to resolve is that they did not all have the same ideals.

Siegel focuses his discussion on the importance of critical reasoning, and argues that that gets short shrift under the economic conception. This may be true, and it may indicate a defect in the FCAT. But it is clearly possible to test critical reasoning abilities via standardized tests. The Law School Admissions Test (LSAT) does it. The exam contains complicated passages followed by subtle questions with cleverly crafted

misleaders. Understanding the passages and reasoning critically are required to figure out what actually follows from, is compatible with, or is supported by the passage. So if the ability to reason critically is an intrinsically and/or extrinsically valuable product of primary or secondary education, there is no reason to suppose that it would be particularly difficult to design standardized tests to see whether it has been acquired.

The core of Siegel's concern is not critical thinking per se. It is that there may be educational goals whose satisfaction cannot be demonstrated by High Stakes Tests. He is surely right about this. It is hard to imagine a standardized test for sensitivity, originality, co-operativeness, or openness to new approaches. In itself, this is not a problem. But to the extent that educators become convinced that the results of High Stakes Tests are *the* measures of achievement, they may become blind to the value of objectives that cannot be measured by such tests. This does not of itself undermine the value of such tests where they are appropriate. But it does argue that they must be construed as having limited value. Even if there is reason to say that passing such a test is a necessary condition for promotion or graduation, unless we are convinced that the test tests for all relevant educational objectives, performing adequately on the tests should not be deemed sufficient.

Siegel's argument thus shows that before we introduce a High Stakes Test we need to determine what we consider to be educationally valuable and why. Then we need to figure out how, if at all, it is possible to test for it rigorously. Without answers to the basic philosophical question – What sort of education is a good education? – whatever we test for is ungrounded.

Schrag is concerned about the achievement gap between races. Even if we

concede with Siegel that the economic argument is not the whole story, the difference between having and lacking a high school diploma has major economic consequences, and the achievement gap between races has the result that a far higher proportion of minority students end up lacking high school diplomas. This is a brute fact. The question though, is what to do about it. As Schrag notes, there is no easy or obvious answer. It is not even clear why having or lacking a high school diploma makes a difference. If the diploma itself is the decisive thing, then we should simply hand them out with wild abandon. Social promotions should be acceptable; sticking it out until the end of 12th grade should be all that it takes. Maybe we should settle for even less. But if the diploma is a mark of some sort of achievement, then the question is harder. What kind of achievement is it, and can we insure that more people have it?

If the diploma marks a competitive advantage in a zero sum game, closing the achievement gap just levels the playing field. In that case, the number of people who are permanently economically disadvantaged remains constant, although the racial profile of that group changes. This is a step in the direction of justice, but plainly no one can be satisfied with this outcome, any more than we can be satisfied with the status quo. If this is our situation, then fundamental economic reforms are morally required to make it possible for more people to lead productive and satisfying lives. But it is not obvious that we are in a zero sum game. Perhaps a better educated workforce would constitute a resource for economic expansion, thus generating more jobs. In that case, the remedy is, to a larger extent, a matter of education, and the role of High Stakes Tests is more clearly relevant. We can worry about the ways the lack of a diploma affects economic well being for members of identifiable groups – minorities, people with limited English, etc. –

or in general. But we need to worry about it. Closing the gap by watering down the achievements of whites or fluent speakers of English would presumably not be a good thing. Whatever the reason for the achievement gap, we need to figure out how to bring up the level of achievement and economic prospects of those at the bottom.

Curren construes the issue in terms of the problem of knowledge of other minds. This strikes me as a brilliant way to look at the matter. Any assessment of student learning involves having reason to believe that our evidence affords knowledge of other minds. This is so for portfolios, term papers, pop quizzes, class discussions, and so forth. Therefore if there is any epistemological justification for assessing students in any way at all, knowledge, or at least epistemically well grounded belief about other minds must be possible. This means we can either throw in the towel, or bracket skeptical possibilities and take the problem to be this: Assuming knowledge or well-grounded belief about the contents of other minds is possible, how can we obtain that knowledge? To answer this requires a theory about what constitutes sufficient evidence for attributing a state of mind to someone, how reliable that evidence is, in what circumstances the evidence is reliable, and so on.

The skeptical possibility is a non-starter for any practical enterprise. It could, of course, be true. But it affords no basis for action. It would certainly undermine any sort of educational endeavor, not just assessment. If we can know nothing about other minds, we cannot feasibly hope to bring it about that other minds (if they exist) change in worthwhile ways. Hence, if we are going to embark on education at all, we need to assume (a) that other people have minds and (b) that it is possible, somehow, to gain knowledge or justified beliefs about the contents of their minds.

Davis[2], according to Curren, espouses a sort of Quine-Davidson holism, which Davis thinks undermines any hope of attributing specific beliefs to students. The reason, very roughly, is that if Quine-Davidson holism is true, there is no such thing as specific belief. Rather, anything that a person might express by saying 'I believe that-*p*' is really a whole theory or belief cluster of which *p* is just one small part. If Quine and Davidson are right, beliefs are not discrete mental representations that exist in isolation from one another, hence not items that we squirrel away one by one.[3] You cannot have the belief that a dynamo produces direct current without having a cluster of other beliefs about electricity. This strikes me as right. Quinean arguments convince me that the point is a general one. But sheer common sense makes the case for beliefs about dynamos. Anything that could count as a belief about a dynamo must be part of a cluster of beliefs about electricity and how to make it.

But the Quine-Davidson argument is *not* a skeptical argument.[4] To say that we cannot believe that dynamos are sources of direct current without believing a lot of other things about dynamos and electric currents is not to say that we cannot believe that dynamos are sources of direct current. We must interpret a statement about dynamos as part of a (perhaps unarticulated) theory about related matters. Multiple theories embed the statement and there is no basis for distinguishing among them. But they are in an important sense equivalent, each being acceptable in the same circumstances as the others. Radical translation and radical interpretation work. Translation and interpretation are indeterminate in the sense that there are multiple correct interpretations.[5] But there is a standard for correct interpretation, and there are ways to determine whether we have met that standard.

Davidson contends that to explain how mental events cause physical events, we need neither appeal to weird causal mechanisms nor assume that mental types are identical with physical types. If my decision to raise my hand is identical with some neural state with the requisite physical powers, then my decision causes my hand to rise. This does not require that such decisions as a type are identical with neural states as a type. All that is needed is that the token of my decision (a psychological token) is identical to some neurological token.[6] On Davidson's view, although each mental event is identical to some physical event, it is not the case that each mental event-type is identical to some physical event-type. The same decision, taken an hour later, might be realized in some different neural structure. There is no hope of reductionism of the mental to the physical, or even of anything but the weakest of supervenience of the mental on the physical.

Davis seems to think that this view vitiates any justified attribution of beliefs. Curren, I think, takes this argument more seriously than it merits. Things with a common functional or dispositional characterization need share no underlying structure. This is so whether they are construed as single-track or multi-track dispositions. So the claim that 'The "practice of identifying" multi-track dispositions, such as magnetism, requires an "underlying" structure "which serves to explain [the] range of phenomena" specific to each disposition' strikes me as just wrong. Rylean examples show this.[7] Consider casting a vote: There is not one physical action that constitutes voting. Sometimes saying 'aye', sometimes raising your hand, filling in a ballot by hand or electronically, writing in a name, flicking a switch in a voting booth, and so on. But we can generally tell whether someone has voted. Plainly certain background conditions have to obtain before saying

'aye' or flicking a switch or writing 'John Kerry' on a piece of paper, constitutes voting. So the act in splendid isolation would not be a vote. But sometimes those conditions obtain, and we know or at least have good reason to believe that they obtain. Hence we know that someone has voted. Multi-track dispositions may more strongly supervene on underlying structures, but I do not see that they must do so in order for us to recognize and correctly ascribe them.

Davis's conviction that beliefs must be shown to strongly supervene on neurological structures in order justify standardized seems simply false. Whether one takes a realist stance toward folk psychological entities is irrelevant. We presumably need a distinction between believing that, say, Fredrick the Great was King of Prussia and not believing it, before we can ask about it on a test. And we presumably need some way to tell whether someone does or does not harbor the belief. But there is no reason to think that only realists about beliefs have the requisite resources. Quine's position is a form of behaviorism. The belief that-*p*, according to Quine, is just a complex propensity to behave. If we know what that propensity is, we can elicit evidence of it -- that is, elicit examples of the behavior. In this case, just asking, 'Was Fredrick the Great King of Prussia?' would probably do the trick.

Everything I have said so far is entirely general. It has nothing special to do with High Stakes Testing. It applies equally to Low Stakes Testing, portfolios, class discussions, term papers, and so forth. So the question is whether holism raises any special problems for High States Testing.

The Davis worry has to do not with the High Stakes, but with the fact that such tests test one thing at a time. Scoring the test, he thinks, is 'interpreting an isolated act of

an unknown actor'. This does make it seem as though the problem of radical interpretation will cause difficulties. If our total evidence about the native was that he uttered 'Gavagai' on a single occasion, we would be hard pressed to figure out what he was talking about. Because his utterance was an isolated act of an unknown actor, we would not have enough data. But it is not clear that properly designed tests are best scored as though they involved interpreting isolated acts of unknown actors.

If the belief that-*p* cannot be divorced from the cluster of beliefs that give it its content, there is no way to test for the belief that-*p* without testing for the entire cluster, since the belief does not have any identity apart from the cluster. This might be true (I think it is), but if it is, it is in a way that is totally innocuous. A good short answer question to investigate what a student believes about dynamos or whatever could implicate a whole cluster of beliefs. That is why, for example, well designed multiple choice tests have misleaders in them. If you did not know enough about electricity, you would opt for the misleader rather than the right answer. So getting the right answer to the dynamo question affords evidence that the student has the requisite cluster of beliefs.

One might worry that a test taker could be in a position to answer the question correctly simply by memorizing a list of facts, without having the background knowledge. So when is her getting the right answer reason to believe she has the knowledge we are interested in? To some extent, it depends on test design. A test can be constructed so as to make it unlikely that one could answer correctly unless one had the requisite background beliefs. It could even set thresholds to block the effects of lucky guesses. In that case students get no credit for the dynamo section unless they get a given proportion of the questions that bear on dynamos right. This just shows that if we are

going to test for beliefs about something, we have to develop reasons to think that the limited amount of information the test reveals constitutes good reason to think that they have the rich cluster of beliefs that we are interested in. The point holds regardless of what method of assessment we use.

There remains a Quinean problem to which, ironically, it turns out that the high stakes of High Stakes Testing supplies an answer. It derives from the interdependence of belief and motivation. In one of his more amusing formulations, Quine says, 'Imagine a dog idling in the foreground, a tree in the middle distance, and a turnip lying on the ground behind the tree. Either of two hypotheses, or a combination of them, may be advanced to explain the dog's inaction with respect to the turnip: Perhaps he is not aware that it is there, and perhaps he does not want a turnip.'[8] If we know what the dog wants, we can figure out what he believes, and if we know what he believes, we can figure out what he wants. But if we don't know either, we are stumped, since multiple ascriptions of belief and desire accord with the evidence.

Assume that students are like the dog. Then the evidence provided by a test reveals belief only if we control for desire. It turns out that making the test High Stakes, does that. Originally the MCAS was supposed to be used to assess schools, not individual students. By seeing how students in a school performed on a statewide test, it would be possible to compare schools with one another. So initially test results were not taken to reflect well or badly on the individual students. As a result, many students flatly refused to take the tests seriously. They made no effort to answer the questions correctly, since no negative consequences accrued to them personally if their answers were wrong.[9] The enterprise of trying to use the tests only to identify poorly performing schools failed

abysmally. Now the tests are High Stakes. A student cannot graduate from a public high school in Massachusetts without passing the 10th grade math and English tests. This has the effect of controlling for desire. We can take it for granted that most students want to graduate, so we can take it that their answers reflect their beliefs.

The assumption that the students are like the dog goes too fast, though. When it comes to language users, Quine believes, there are three interdependent variables – belief, desire, and meaning – rather than two.[10] So by controlling for desire, a High Stakes Test enables us to test for the fusion of belief and meaning. It might seem that this still leaves us in an intolerable situation. In fact, it does not. Multiple attributions of belief and meaning yield the same outputs in terms of behavior and dispositions to behave. These outputs may be all we care about. In that case, discriminating between belief and meaning is unnecessary. But if we want to get any more fine-grained information about beliefs, we need to find a way to control for meaning as well. Although meaning is a tricky concept, there is no reason to think we could not control for it, at least as far as is necessary for these purposes.

Scorers need not interpret the tests as consisting of a series of isolated acts of unknown actors. Even if they do not know, and should not know, the particular students whose tests they are grading, there are lots of things they can know or reasonably presuppose about the test taker. She is, say, a tenth grader in Massachusetts, who is supposed to be acquainted with a given curriculum and is aware that she is supposed to be acquainted with that curriculum. She is, and knows that she is, supposed to be reading the questions as written in standard academic English and writing her answers in an appropriate format. We know rather a lot about the population of students who are taking

the test, and about the sorts of things they are apt to think about how to take tests of this kind. Grading an MCAS exam is not like trying to decipher the Rosetta stone. Moreover, if it seems worth doing, we can incorporate internal checks into the tests to increase the probability that the students interpret the questions in the ways the test givers intend.

This brings us to the final part of Curren's paper. The conviction that you cannot achieve sufficient inter-rater reliability for essay tests, and the conviction that you cannot test for fine-grained knowledge without essay tests are both unfounded. Curren mentions several reasons for this. A further reason is that the tests in question only purport to measure thresholds. They are not anything like absolute measures of knowledge. It may be that a standardized test can only measure a proxy for a target outcome, or an element of it. But whether the student has the proxy ability or the element of the outcome can be important, if we have good reason to think that anyone who displays the tested abilities or knowledge does to a suitable degree achieve the outcome. So one reason why the tests may be informative is precisely because the have relatively low standards. Acing the math MCAS does not show that a student is a brilliant mathematician, but it is very good evidence that he has solid grounding in high school mathematics, which after all, is all that the test purports to show.

The conclusion is that Quine-Davidson holism does not supply an in-principle reason for objecting to standardized tests, even High Stakes standardized Tests. It does, however, provide reason to think that the tests have to be well designed, so that the responses the test takers give are reliable indicators of their overall mastery of the subject. This is a general point about assessment though, not specific to High Stakes

Tests, standardized tests, standardized short-answer tests, or any other sort of tests.

One final point: If I am right that High Stakes Tests control for desire and thereby disclose belief – if, that is, when the stakes are high enough, most students try to pass – then Schrag's concern with justice reemerges. It may not be unreasonable to lay the blame for their own failures on a few students who did not adequately apply themselves. But it is not plausible to hold large numbers of students, conveniently clustered in particular (often underfunded, largely minority) schools, responsible for their own failures. So schools with high failure rates should be deemed *to have failed* their students. They have not equipped the students with the knowledge and abilities needed to pass the tests. And if such knowledge and abilities are genuinely indicative of being adequately educated, the failure rate is evidence that they have not adequately educated their students. If such schools lack the resources to do better, society has failed both the schools and the students. This is an issue of justice. If 'ought' implies 'can', we owe it to students to give them the resources they need to pass the tests that we say they ought to pass. If justice is fairness, schools with high failure rates, and the political institutions that sustain such schools, do the students attending them an injustice.

Notes

[1] Harvey Siegel, 'High Stakes Testing, Educational Aims and Ideals, and Responsible Assessment'; Randall Curren, 'Educational Measurement and Knowledge of Other Minds'; Francis Schrag, 'High Stakes Testing and Distributive Justice'; *This Journal*. All three papers were presented at the symposium on 'School Accountability and High Stakes Testing' sponsored by the Association for Philosophy of Education, in December 2003. I am grateful to these authors, to Nell Noddings and members of the audience for useful suggestions.

[2] Andrew Davis, *The Limits of Educational Assessment*, Oxford: Blackwell, 1998.

[3] See W.V.Quine, 'Two Dogmas of Empiricism,' *From a Logical Point of View,* New York: Harper Torchbooks, 1961, pp.20-46; Donald Davidson, 'Radical Interpretation,' *Inquiries into Truth and Interpretation,* Oxford: Clarendon, pp. 125-139.

[4] See my *Considered Judgment*, Princeton: Princeton University Press, 1996, pp. 208-220.

[5] W.V. Quine, *Word and Object*, Cambridge: MIT Press, 1960, pp. 26-80.

[6] Donald Davidson, 'Mental Events,' *Essays on Action and Events*, Oxford: Clarendon, 1980, pp. 207-225.

[7] See Gilbert Ryle, *The Concept of Mind,* New York: Barnes & Noble, 1949.

[8] W.V. Quine, 'On the Nature of Moral Values,' *Theories and Things,* Cambridge: Harvard University Press, 1981, p. 55.

[9] By 'no effort', I do not mean merely 'did not try very hard'. We are dealing with middle school and high school students here. Many of the answers were sarcastic, intentionally wrong, random, and so forth.

[10] Quine, *Word and Object*, pp. 73-80.

Biographical Note: CATHERINE Z. ELGIN is Professor of the Philosophy of Education at Hairvard Graduate School of Education. She is the author of *Considered Judgment*, (1996), *Between the Absolute and the Arbitrary* (1997), *With Reference to Reference*, (1983), and co-author with Nelson Goodman of *Reconceptions in Philosophy and Other Arts and Sciences,* (1988). Her work has been supported by the National Endowment for the Humanities, the Mary Ingram Bunting Institute of Radcliffe College, the John Dewey Society, the American Council of Learned Societies, and the Andrew W. Mellon Foundation.