TESTING

By ***Banesh Hoffmann***

Trained as a mathematical physicist, Prof. Hoffmann has served since 1937 as a member of the Department of Mathematics at Queens College in New York.

IF scholastic aptitude and ability could be objectively determined by measuring people's heights, there would be no need for this article. But aptitude and ability cannot be measured so simply. Their assessment is at best tentative and fallible.

Nowadays, though, with testing largely mechanized, this tends to be forgotten. The number of students to be tested is rising rapidly, and with it the temptation to place ever greater reliance on multiple-choice tests. There are machines that can grade 6000 multiple-choice answer sheets per hour, and numerical grades emerging from such machines convey an impression of gleaming precision. Yet the value of these grades clearly depends on the quality of the tests—and, indeed, on the very nature of the tests, for there are inherent limitations in the multiple-choice format.

Many things can be said in favor of multiple-choice tests; and the makers of these tests can be relied upon to say them. Much, too, can be said against these tests, as critics are by no means reluctant to point out. But one of the alarming aspects of the present situation is the degree to which the test makers have already taken over effective control of testing. Can the trend be reversed, or is it already too late?

How good are the testers? And how good are the multiple-choice tests on which they place so much reliance? If there is reason for doubt as to the quality of the tests and the level of competence of the test makers, we must be vitally concerned, not only as scientists but as responsible members of the community of scholars.

Here is a question in physics, made by the Educational Testing Service and published in 1954 by the College Entrance Examination Board in a booklet entitled *Science*, describing the College Board's Science Achievement Tests. It is labeled "difficult".

> 64. A ray of white light is broken up into a spectrum by a prism of colorless glass because
>
> (*A*) the angle of incidence exceeds the critical angle of glass
>
> (*B*) white light is a composite of many frequencies
>
> (*C*) the glass absorbs certain frequencies and not others
>
> (*D*) the amount of refraction differs for light of different wave lengths
>
> (*E*) white light from an incandescent solid produces a continuous spectrum.

I have tried this question on several scientists. Some pick answer *D* at once, and that happens to be the wanted answer. But others are hard put to it to choose between answers *B* and *D*, realizing that each gives only a partial reason and that what is needed is a combination of both. Moreover, when this is pointed out to those who quickly picked *D* they usually join the band of waverers.[1] Of the two partial reasons in answers *B* and *D*, which one is the more fundamental? Surely Newton's magnificent discovery of the composite nature of white light. The two ideas in *B* and *D* are linked together, of course; but the controversy that followed Newton's discovery was over the prime question of the nature of white light rather than over the details of the refraction of light. Thus a deep student would be justified in preferring *B* to *D*. But the crux of the matter is that the test maker has used ambiguity as a substitute for genuine difficulty—as test makers frequently do—and has thus produced a question that, in the case of good students, does not measure understanding of science so much as understanding of the workings of the mind of the test maker.

THE matter of testing has many ramifications, of which the existence of defective multiple-choice questions is by no means the most important. But, as I have explained more fully elsewhere,[2] the manner in which test makers treat *general* criticisms of current test procedures led me to try a new, more specific approach. It has now produced evidence regarding the quality of test makers that scholars may well find disturbing.

Briefly, the new idea was to exhibit defective sample questions and to challenge the test makers to defend them explicitly. Many years and much maneuvering were needed [3] to induce the Educational Testing Service to offer a public defense of specific sample questions. The defense that it now gives proves to be a revealing document.[4] The present article discusses the part of this document that pertains to two sample questions in science that were exhibited in *Harper's*. I hope to discuss the rest of the document elsewhere.

---

[1] One person to whom I showed this question argued that *D* was superior to *B* because it contained *B* by implication, "light of different wave lengths" referring to the "white light" mentioned in the stem of the question. If ETS wishes to argue along these lines, I am ready to discuss the faulty English of answer *D* and show that the implication is not valid.

[2] Banesh Hoffmann, "The Tyranny of Multiple-Choice Tests", *Harper's Magazine*, March, 1961, p. 37. This article will here be referred to as "*Harper's*".

[3] See, for example, (a) Banesh Hoffmann, " 'Best Answers' or Better Minds", *American Scholar*, 1959, p. 195, and (b) the published correspondence regarding (a) between Henry Chauncey, President of the Educational Testing Service, and myself, ibid. p. 538–9. Item (b) was preceded by more extensive, unpublished correspondence between members of ETS and myself.

[4] "Explanation of Multiple-Choice Tests", April, 1961, obtainable on request, and without charge, from Educational Testing Service, Princeton, N. J.

Here is a question in chemistry, made by the Educational Testing Service. It appears in the same College Board booklet, *Science,* from which the previous question was taken.

> 54. The burning of gasoline in an automobile cylinder involves all of the following *except*
>
> (*A*) reduction (*B*) decomposition (*C*) an exothermic reaction (*D*) oxidation (*E*) conversion of matter to energy.

I exhibited this question in *Harper's* and pointed out that "the average chemistry student quickly picks the wanted answer *E* . . . but the student who is unfortunate enough to understand, even if only in an elementary way, what $E = mc^2$ is really about finds himself at a distinct disadvantage" because he knows that all of the released energy comes from the conversion of matter to energy and so correctly concludes that no answer is correct.

The Educational Testing Service [4] defends this question as follows.

"*Explanation*—The superior student is as aware of the *classical* concepts of matter and chemical change as he is of the model of *modern* physics. He is likely to be more aware than is the average student that the 'conversion of matter into energy' has been demonstrated only for nuclear changes. Perhaps he realizes that if the energy freed by the burning of gasoline comes from the conversion of mass to energy, the loss in mass is only about a ten-billionth of the mass of the gasoline burned, a loss too small to be measured by available methods.

"When such a student is faced with the above question, he should realize that the classical concepts of matter and chemical change provide the framework in which the question is asked. He also recognizes that the first four processes listed are obviously and immediately involved in the burning of gasoline, and he selects response *E* as the required answer."

Note the curious implication of the words I italicize in this passage: ". . . the 'conversion of matter into energy' has been demonstrated *only* for nuclear changes. Perhaps he realizes that *if* the energy freed by the burning of gasoline comes from the conversion of mass to energy. . . ." Does one not receive the impression that, in order to defend its question, ETS is prepared, if necessary, to abandon $E = mc^2$?

The remark that the loss of mass is "too small to be measured by available methods" [5] is hardly relevant to the crucial question of whether mass is or is not converted into energy in the burning of gasoline. Can ETS produce a competent physicist or chemist who would risk his reputation by denying in public that, according to current concepts, *all* of the released energy comes from the conversion of rest mass? If *all* the released energy comes from this conversion, the process is not a negligible one here.

ETS says that the superior student "should realize that the classical concepts of matter and chemical change provide the framework in which the question is asked". It is worth remarking in this connection that $E = mc^2$ is over fifty years old. Why should the superior student realize that he is to use only the "classical" concepts that ETS has in mind? Does he not see a non-"classical" relativistic answer among the choices? Was this answer put there deliberately, or was ETS, at the time it framed the question, unaware of the meaning of $E = mc^2$? Note how damaging are the implications if we do assume that ETS was fully aware of the meaning of $E = mc^2$ and deliberately included answer *E* nevertheless. For we must then ask: what was its motive in doing so? To make a question with no correct answers? Let us hope not. Then what? To penalize the superior student? One doubts that ETS would say so; yet the question is surely easier for the student who does not understand $E = mc^2$ than for the student who does. Is the latter student supposed to compensate for the deficiencies of the test maker by reading possibly hazardous amendments into the question as worded? That way lies chaos—not "objectivity". If the superior student does decide to pick answer *E,* does he not do so with contempt for the test maker, and with cynical disregard of scientific facts? Should he be rewarded for his willingness thus to place expediency above scientific integrity? If tests are training students to respond in this way, are they not having a deleterious effect on education? Perhaps, after all, it is more charitable to assume that ETS was ignorant of the meaning of $E = mc^2$, even if this does imply a certain lack of candor on its part now.

Here is the second of the two science questions that I discussed in *Harper's.* It, too, comes from the College Board booklet, *Science.*

> 65. Potassium metal loses electrons when struck by light (the photoelectric effect) more readily than lithium metal because
>
> (*A*) the potassium atom contains more protons than does that of lithium
> (*B*) the valence electron of potassium is farther from the nucleus than is that of lithium
> (*C*) potassium occurs above lithium in the electrochemical series
> (*D*) the potassium atom contains more electrons than does that of lithium
> (*E*) the potassium nucleus is larger than that of lithium.

Discussing this question in *Harper's,* I wrote: "The wanted answer is *B*. Let us accept it as a factually correct answer and ask whether it is the best answer. . . . a well-prepared and inquiring student . . . may say to himself, the sentence in question can be plausibly completed with the statement that the 'valence electron of potassium is farther from the nucleus than that of lithium'. But he then sees that answer *D* accurately (if ungrammatically) states the reason *why* this is so. . . . Thus, *D* is a correct answer too. And *D* is a more profound answer than *B*.

"But our student is not finished, for he realizes that the reason why there are more electrons in the potas-

[5] By weighing, that is; for it can certainly be measured by determining the amount of energy released and dividing by c^2.

sium atom than in the lithium atom is to be found in answer *A:* the atom of potassium 'contains more protons than does that of lithium'. Thus, if *D* is a correct answer, so is *A*. And *A* cuts deeper than *D*.

"Finally he hesitates to dismiss *E*, knowing that the nucleus of potassium 'is larger than that of lithium' because it contains more neutrons and protons. Thus if *A* is a correct answer, so also is *E*.

"In view of the above, most of us would agree with the College Board that the question is 'difficult'. But with us this is merely a matter of opinion. With the test experts it is an objective, scientific, no-nonsense fact based on statistics. Of course, the statistics do not reveal that the wording of the question is vague. Nor that the examiners have chosen the most immediate and superficial answer, thus penalizing the candidates with more probing minds, as they so often do."

The Educational Testing Service defends this question as follows.[4]

"*Explanation*—The technical terms must be considered in studying this question. The photoelectric effect is exhibited by an element if, in atoms of the element, an electron is so loosely bound that visible light provides enough energy to free that electron from its atom. Since electrons are negatively charged, most of them are too strongly attracted to the positively charged atomic nucleus to be freed by light. The farther from the nucleus an electron is found, the more likely it is that light will be able to free the electron and that the photoelectric effect will be observed.

"Since the outer—or valence—electron of a potassium atom, on the average, is farther from the nucleus than the valence electron of a lithium atom, of these two the element that shows the photoelectric effect is potassium. Response *B*, the accepted response to this question, is based on this reasoning.

"Dr. Hoffmann agrees to accept response *B* and then begins to study other responses to see whether they can account for *B*. He reasons that if *B* is the cause of the photoelectric effect and if *D* is the cause of *B*, then *D* must be the cause of this effect.

"Cause-effect relations in science are difficult to reduce to the confines of one response to a multiple-choice question; to find a chain of causes, such as Dr. Hoffmann proposes, in a single question would be most surprising, but we must look.

"It is quite true that if one limits his consideration to a family of elements, like the one that contains potassium and lithium, the greater the number of electrons in an atom, the farther from the nucleus is the outer electron likely to be found. Is there a cause-effect relation here? Potassium has 19 electrons, calcium has 20; yet the outer electron of the calcium atom, on the average, is closer to the nucleus than is the valence electron of potassium. Indeed, of the elements whose atoms have progressively more electrons than potassium, krypton, with 36 electrons, is the first element for whose atoms the outer electron is normally farther from the nucleus than is the valence electron of a potassium atom. A larger number of electrons in an atom clearly does not 'cause' the outer electron of an atom to be farther from its atomic nucleus. If *D* does not 'cause' *B*, it can hardly be said to 'cause' the photoelectric effect. The other responses cited by Dr. Hoffmann as 'causes' of the effect can be criticized in the same fashion."

This is a remarkable defense. There is no escaping the conclusion that ETS believes not only that the photoelectric effect is confined to visible light, but also that it is not exhibited by lithium.[6]

Perhaps because of these initial misapprehensions, ETS does not even understand what its own question is about. It clearly believes that the question asks for the cause of the photoelectric effect.[7] Actually the question asks why "potassium metal loses electrons . . . *more readily* than lithium metal", the parenthetical phrase "the photoelectric effect" being in fact relevant to lithium as well as to potassium. Does ETS expect the superior student to read possibly hazardous amendments into this question too? How can the superior student hope to guess what is in the examiner's mind when the examiner makes so many unpredictable elementary blunders?

Despite what ETS says, the wanted answer, *B*, does not give the cause of the photoelectric effect.

Is *B* a good answer to the question as worded? The crucial quantity is not distance from the nucleus but amount of energy needed to remove the electron from the metal, and this depends on the state of the metal. Let us, for the sake of argument, assume, as ETS seems to do, that we are dealing with individual atoms, as in the gaseous state. Then the crucial quantity is the ionization potential, which is not related in a simple way to the distance of the electron from the nucleus. There *is*, however, a regularity when one confines oneself, for example, to elements of the alkali-metal family. Since the question pertains specifically to lithium and potassium,[8] one can, for the gaseous state, make a plausible case for the relevance of answer *B*—not as a direct answer, but as a link in a causal chain of which a more immediate link would involve the values of the ionization potentials.

But ETS wishes to argue against other answers. So it enlarges the scope of the question by allowing several elements to enter. Note how, in so doing, it plays fast and lose with the phrase "the valence electron", replacing it by the phrase "the outer electron" wherever it would draw attention to the nature of the maneuver. Using this enlargement of scope as an argument against

[6] For example, it says "the photoelectric effect is exhibited . . . if . . . an electron is so loosely bound that visible light provides enough energy to free that electron . . .", and, later, ". . . of these two [potassium and lithium] the element that shows the photoelectric effect is potassium."

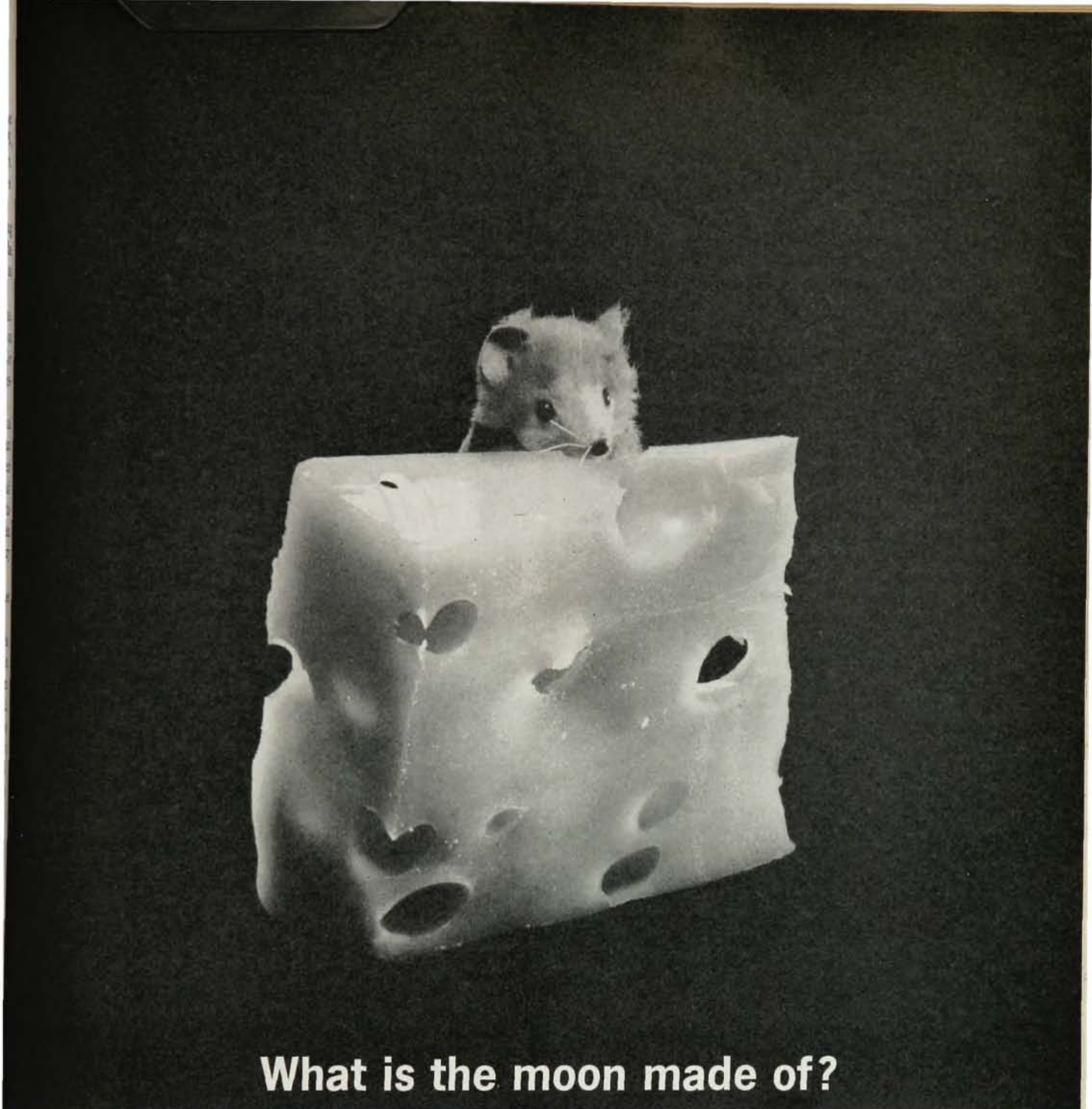
[7] Note, for example, the context of the sentence "Response *B*, the accepted response . . . is based on this reasoning". Note, too, the words:

"He reasons that if *B* is the cause of the photoelectric effect and if *D* is the cause of *B*, then *D* must be the cause of this effect,"

"If *D* does not 'cause' *B*, it can hardly be said to 'cause' the photoelectric effect", and

"The other responses cited by Dr. Hoffmann as 'causes' of the effect . . . ."

[8] This is clear from the repeated references to these elements in the question and the answers, not to mention the implications of the phrase "the valence electron" in the wanted answer, *B*.



What is the moon made of?

No. Guess again.

Potassium, uranium, thorium? Closer, but still guesswork. And guesses they'll be until man puts scientific instruments on the Moon to gather surface and sub-surface data and transmit these data to Earth.

The right answers will come with unmanned lunar spacecraft projects directed by Caltech's Jet Propulsion Laboratory for the National Aeronautics and Space Administration.

The planned Lunar Exploration Program begins with JPL's Ranger Project that will soon hard-land 50-pound instrument packages on the Moon to measure Moon quakes and temperature and radio their findings back to Earth.

Following the Ranger, the Surveyor will *soft*-land several hundred pounds of sensitive instruments on the Moon. Its objectives are to measure the physical properties of the Moon and analyze the composition of surface and sub-surface samples. Knowledge from these projects is essential to eventual manned landings on the Moon.

Under JPL direction, unmanned spacecraft for these projects and probes to the planets are being designed. Many disciplines are involved. Physics, electronic engineering, metallurgy . . . it's a long list.

It's a big job. To do it right, JPL must have the best technical people in the country. People who want to know . . . who want to be part of the greatest experiment of mankind. If you're that kind of people, JPL is your kind of place. Write us today.

JET PROPULSION LABORATORY
4810 Oak Grove Drive, Pasadena, California
Operated by California Institute of Technology for the National Aeronautics and Space Administration

jpl

All qualified applicants will receive consideration for employment without regard to race, creed or national origin / U.S. citizenship or current security clearance required.

answer *D*, the ETS triumphantly points out that "of the elements whose atoms have progressively more electrons than potassium, krypton, with 36 electrons, is the first element for whose atoms the outer electron is normally farther from the nucleus than is the valence electron of a potassium atom".

But the triumph proves illusory, for the argument boomerangs. If ETS accepts answer *B*, how does it propose to account for the fact that krypton does *not* "lose electrons when struck by light more readily than" potassium?

ETS can not have it both ways. To accept *B* it must limit the number of elements involved, in which case it may not deny that *D* causes *B*. In denying that *D* causes *B* it allows many types of elements to enter; but in this case *B* itself becomes unacceptable and no answer is valid.

When I criticized a different question made by ETS,[8a] the president of ETS, without then defending the question, made various reassuring remarks,[8b] among them that "hundreds of outstanding teachers from schools and colleges work with Educational Testing Service each year to make the examinations we give as good as possible."

IF the leading educational test-making organization, aided by hundreds of outstanding teachers from schools and colleges, not only comes up with sample questions as vulnerable as these two, but, after saying that questions made by ETS are "as good as possible", can offer no better defense of these questions than that exhibited here, what shall we think of the quality of its tests—not to mention the true nature of the much-vaunted "objectivity" of these tests? I hope the reader will understand, and ETS too, that I discuss ETS here for a reason that is complimentary to that organization: one makes the strongest case by criticizing the best test makers, not the worst. In connection with this matter of the quality of the best organizations, it may be worth mentioning that on a panel discussion on NBC-TV, in which Mr. Chauncey and I participated, Mr. Chauncey said, of the second and third of the three College Board science questions above, that [9] "the College Board Committee on chemistry met [in mid-March, 1961] and looked over those questions again, and they seemed to be entirely satisfactory, good questions."

Anticipating the argument that I have to seek far and wide to glean a few isolated defective questions, I prepared, a few years ago, a list of twelve challenge questions—representing five percent of the sample questions in two College Board booklets. As I pointed out [see ref. 3b], five of these twelve questions are of one particular kind, and they constitute 24 percent of the supply of that kind in the booklets. Four of the twelve questions have been previously exhibited, and they have now been defended by ETS.[4] The "gasoline" and "potassium" questions above belong to this category. The "spectrum" question exhibited above is another of the twelve challenge questions for ETS to defend. There were only three sample questions in the College Board booklet, *Science*, dealing with "ability to interpret cause and effect relationships". One of these was said to be of average difficulty. The other two were said to be difficult—they are the "potassium" and "spectrum" questions above.

My object in exhibiting defective sample questions, and in noting here the quality of the defense they elicit from a leading test maker, is not to present the whole case for and against current test procedures, but merely to make a prima facie case for the setting up of a distinguished committee of inquiry to look into the whole matter of testing. If all is well, ETS and the College Board have nothing to fear from an inquiry. They should welcome one. Indeed, on the "Open Mind" panel discussion,[10] Mr. Chauncey implied that a committee of inquiry might well be a good thing. In the long run, ETS would probably have more to gain than lose from the inquiry, even if the multiple-choice test should emerge with its repute somewhat diminished. For the high position of ETS among testing organizations would doubtless be confirmed; and, as a responsible organization, it would certainly be among the first to put significant recommendations of the committee into practice. It might well be the organization chosen by the committee to conduct certain experiments designed to illuminate aspects of the testing problem.

Judging by the overwhelmingly favorable mail that I have received about *Harper's*, there is strong feeling among educators against the improper use of multiple-choice tests. If products of the best test-makers are of questionable quality, what of multiple-choice tests made by lesser organizations? And what of multiple-choice tests made by individual teachers for their own use? The minimum concern of the committee of inquiry should be the problem of policing tests to ensure that they meet high standards. In this connection, the qualifying tests used by the National Merit Scholarship Corporation deserve careful scrutiny, for they are used *in order to exclude some 98 percent of the candidates from further consideration*.[11]

If the arguments used by ETS to defend the above science questions are a fair sample, may we not wonder about the merits of other arguments, made by it and other leading test makers, in support of their products and of multiple-choice tests in general? As scientists, we know the importance of educational procedures that stimulate rather than discourage precision, depth, and creativity. Dare we ignore the effect of tests on education?

[9] Transcript of THE OPEN MIND, April 2, 1961, NBC Television "How Good is Educational Testing?" See page 4.

[10] See reference 9, pages 17–19.

[11] See the discussion of these tests in *Harper's*, and the letter from NMSC about it in the May 1961 issue of *Harper's Magazine*, where, in its only apparent response to this point, NMSC makes the correct but misleading statement, "We use two tests, prepared by different testing agencies. But we also use students' school records, records of accomplishments outside the classroom, and the judgments of school officials. Finally, all of this information is evaluated not by a machine (although some persons believe it should be), but by experienced, skilled educators who make the actual selection of the National Merit Scholars."