© Washington Post; Reprinted by permission of the D.C. Public Library

How They Write the SAT

By David Owen

Standardized multiple-choice tests, such as the Scholastic Aptitude Test, are more than hurdles on the way to college. The tests have become a pervasive measure of worthiness in our society—even a status symbol, as in, "*My* boy scored double 700s."

The Educational Testing Service, which produces the SAT, encourages this attitude. The company's literature conjures up the image of a testing instrument endowed with the learning and precision of white-coated physicists measuring a rocket's lift-off power. But as should be apparent to anyone who has taken these tests, the white-coated image is just that. Many of the test questions are ambiguous, arbitrary, and downright silly.

The principal difference between the SAT and a test that cannot be graded by machine is that the SAT leaves no room for more than one correct answer. It leaves no room, in other words, for people who don't see eye-to-eye with ETS. Understanding how the test-makers think is one of the keys both to doing well on ETS tests and to penetrating the mystique in which the company cloaks its work. Despite ETS's claims to the contrary, its tests are written by people who tend to think in certain predictable ways. The easiest way to see this is to look at the tests themselves.

David Owen is a New York writer. This article is adapted from his book, None of the Above: Behind the Myth of Scholastic Aptitude, *to be published in May by Houghton Mifflin.* © *1985 by David Owen.*

Which test doesn't really matter. Here's an item from a recent Achievement Test in French:

2. Un client est assis dans un restaurant chic. Le garcon maladroit lui renverse le potage sur les genoux. Le client s'exclame:

(A) Vous ne pourriez pas faire attention, non?
(B) La soupe est delicieuse!
(C) Quel beau service de table!
(D) Je voudrais une cuillere!

My French is vestigial at best. But with the help of my wife I made this out as follows:

2. A customer is seated in a fancy restaurant. The clumsy waiter spills soup in his lap. The customer exclaims:

(A) You could not pay attention, no?
(B) The soup is delicious!
(C) What good service!
(D) I would like a spoon!

Now, (B), (C), and (D) strike me as nice, funny, sarcastic responses that come very close to being the sort of remark I would make in the situation described. A spoon, waiter, for the soup in my lap! But, of course, in taking a test like this, the student has to suppress his sometimes powerful urge to respond according to his own sense of what is right. He has to remember that the "best" answer—which is what ETS always asks for, even on math and science tests—isn't necessarily a good answer, or even a correct one. He has to realize that the ETS answer will be something drab, humorless, and plodding—something very like (A), as indeed it is. Thus bright students sometimes have trouble on ETS tests, because they see possibilities that ETS's question-writers missed. The advice traditionally given to such students is to take the test quickly and without thinking too hard.

Medicine freaks

Exactly how does ETS come to write questions like this? ETS is very secretive about its methods. The company has always insisted that its work is too complex and too important to accommodate the scrutiny of outsiders. But with some determined digging around, even an outsider can get an idea of what goes on inside ETS's test development office.

A variety of people write questions for the SAT: company employees, freelancers, even student interns. An "assembler" oversees the process, and once the test is completed, this person gives it to two or three colleagues for a review. ETS's test reviews aren't meant to be seen by the public. The words SECURE and E.T.S. CONFIDENTIAL are stamped in red ink at the top of every page. But I obtained a copy of the review materials for the SAT administered in May 1982, which were used as evidence in a court case.

An ETS test review doesn't take long. The reviewer simply answers each item, marking his choices on ordinary lined paper and handwriting comments on items he feels need improvement. For example, the fourth item in the first section was an antonym problem; students were supposed to select the lettered choice that is the nearest opposite in meaning to the word in capital letters:

4. BYPASS: (A) enlarge
(B) advance (C) copy
(D) throw away (E) go through

The first reviewer, identified only as "JW," suggested substituting the word "clog" for one of the incorrect choices (called "distractors" in testing jargon), because "perhaps clog would tempt the medicine freaks." In other words, if the item were worded a little differently, more future physicians might be tempted to answer it incorrectly. The assembler, Ed Curley, decided not to follow JW's suggestions, but the comment is revealing of the level at which ETS analyzes its tests.

In ETS test reviews, the emphasis is not always on whether keyed answers are good or absolutely correct, but on whether they can be defended in the event that someone later complains. When the second test reviewer, Pamela Cruise, wondered whether answering one difficult item required "outside knowledge," Ed Curley responded: "We must draw the line somewhere but I gave item to Sandy; she could not key—none of the terms were familiar to her. She feels that if sentence is from a legit source, we could defend."

"The legitimate source" tends to be either the American Heritage or Webster's dictionary, depending on which supports the answer ETS has selected. In reviewing item 44, JW wrote, "Looked fine to me but AH Dict. would suggest that *matriarchy* is a social system & *matriarchate* a state (& a gov't system). Check Webster." Curley did this and responded, "1st meaning of 'matriarchy' in Webs. is 'matriarchate' so item is fine." To a criticism of another item he responded, "I had some pause over this, too, but tight by dictionary."

I'd always thought that ETS item-writers must depend heavily on dictionaries. The diction in SAT questions is sometimes slightly off in a way that suggests the item writers are testing words

ETS 'assemblers' don't like to be challenged on their test questions. 'Poop on you,' one responded to an in-house reviewer.

they don't actually use. ("It ain't often you see CONVOKE!" noted JW of a word tested in one item.) SAT items also often test the third, fourth, or fifth meanings of otherwise common words, which can create confusion. In the following item from the same test, the word "decline" is used peculiarly:

17. He is an unbeliever, but he is broad-minded enough to decline the mysteries of religion without ---- them.
(A) denouncing (B) understanding (C) praising
(D) doubting (E) studying

My *Webster's Seventh New Collegiate Dictionary* gives the fourth meaning of decline as "to refuse to accept." This is more or less what ETS wants to say. But the dictionary goes on to explain that decline in this sense "implies courteous refusal esp. of offers and invitations." This usage, and not ETS's, is the proper one. What ETS really wants here is a word like "reject." Cruise made a similar comment in her review, but the item was not changed ("Sounds fine to me and is supported by dictionary," wrote Curley).

Tough enuf

The most important statistic that ETS derives from pretests in terms of building new SATs, is called "delta." Like virtually all ETS statistics, delta sounds more sophisticated than it is. It's really just a fancy way of expressing the percentage of students who consider a particular item but either omit it or get it wrong. (Or, as ETS inimitably describes it, delta "is the normal deviate of the point above which lies the proportion of the area under the curve equal to the proportion of correct responses to the item.") For all practical purposes, the SAT delta scale runs from about 5.0 to about 19.0. An item that very few students get right might have a delta of 16.8; one that many get right might have a delta of 6.3.

ETS calls delta a measure of "difficulty," but this definition is circular. A question is hard if few people answer it correctly, easy if many do. But since delta refers to no standard beyond the item itself, it makes no distinction between one body of subject matter and another. Nor does it distinguish between knowledge and good luck. Delta can say only that a question was answered correctly by the exact percentage of people who answered it correctly. It takes a simple piece of known information and restates it in a way that makes it seem pregnant with new significance.

ETS is almost always reluctant to change the wording of test items, or even the order of distractors, because small changes can make big differences in statistics. Substituting "reject" for "decline" in the above could have made the item easier to answer, thus lowering its delta and throwing off the test specifications. (ETS doesn't pursue the implications. If correcting the wording of a question changes the way it performs on the test, then some of the people now getting it wrong—or right, as the case may be—are doing so *only because* the question is badly written.)

Making even a slight alteration in an item can necessitate a new pretest (or trial run in ungraded portions of existing tests), which is expensive. Revisions are made only grudgingly, even if assembler and reviewer agree that something is wrong. "Key a bit off, but okay," Cruise wrote in regard to one item. JW commented on another: "At pretest I would have urged another compound word or unusual distractor. However it's tough enuf as is."

Test assemblers don't like being criticized by test reviewers. When Cruise described item 26 as a "weak question—trivial," Curley responded in the margin "Poop on you!" The question stayed in the test. Curley's most frequent remark is a mildly petulant "but OK as is," which is scribbled after most criticisms. Assemblers invest a great deal of ego in their tests, and they don't like to be challenged. Sometimes the reviewer is nearly apologetic. "Strictly speaking (too strictly prob-

ably), doesn't the phoenix symbolize death and rebirth rather than immortality? Item's OK, really. It needs Scotch tape." JW concluded this comment by drawing a little smiling face. (The item was not changed.)

The phoenix item, an analogy problem, also drew a comment from Cruise. "Well—item OK—but this reminds me of the kind of thing we used to test but don't do much now—relates to outside knowledge—myth, lit., etc. . . .This might be an item the critics pick on." In ETS analogies, students are given a pair of words and asked to select another pair "that *best* expresses a relationship similar to that expressed in the original pair." The item:

42. PHOENIX:IMMORTALITY::
    (A) unicorn:cowardice
    (B) sphinx:mystery
    (C) salamander:speed
    (D) ogre:wisdom
    (E) chimera:stability

Cruise said she would "be more inclined to defend this item if it were a delta 15." The item had been rated somewhat lower (i.e. "easier"), at delta 13.2. What Cruise *thought* she was saying

The Aptitude Zoo

"You will *not* be admitted to the test center without positive ID," say the instructions ETS and the College Board give students when they sign up to take the SAT. Test scores would be meaningless if college admissions officers couldn't be certain they had been obtained under secure conditions.

Yet when I took the SAT at Julia Richman High School in New York in December 1983, no one asked for my identification. Indeed, when I took out my wallet to get my driver's license, the proctor told me to put it away. She told all the students to put their identification away. "I just need to see your ticket," she said.

Our proctor, who was wearing a jaunty scarf made of blue plastic netting, talked to herself as she waited for students to arrive. She said she was going to give us our test booklets ahead of time but asked that we not open them "in case Dennis walks in." Of course, several students opened their booklets immediately.

Shortly after she gave us our booklets, she told us to begin. We would have 30 minutes to complete the first section, she said, starting now. Then, after we had started, she told us to be certain to fill in the identifying information on both sides of the answer sheet and on the back of the test booklet. Students have to provide quite a lot of information and doing so takes a long time. Students are supposed to do it *before* the test begins, with the proctor leading them through every step; those in my room were cheated out of at least a third of the time allowed for the first section of the test. Even students who ignored the proctor and began working on the test were penalized, because she talked continually.

There was no clock in our room, so the proctor periodically marked the time on the blackboard. Her timing was very approximate, according to my watch. She shaved off a few minutes on some sections, added a few minutes on others.

My desk was so covered with graffiti that the ink from the tabletop sometimes rubbed off on my answer sheet. Erasing these marks was difficult. All the desks had been carved and gouged. It was possible to tear an answer sheet simply by marking an answer.

Proctors are explicitly required to give students a five-to-ten minute break at the end of each hour. Our proctor gave us just one break, very late in the test, only because a student complained. Several students continued working during the break. This, of course, is against the rules. It is cheating. The proctor said nothing, although she was clearly able to see what was going on. Other students worked on sections other than the one they were supposed to be working on. This, too, is cheating.

Students who finished the last section early were allowed to leave. This is absolutely forbidden. It is also extremely distracting to the students still working. The students who left early rustled their coats and papers and talked in normal voices. The proctor talked, too. She stood in the doorway and talked to a woman in the hall. I was working on the last few problems in my second math section at the time. Every time the proctor or one of the students started talking, I lost track of what I was doing and had to begin again. When time finally was called, the proctor allowed the remaining students to continue working on the test. Several still were working when I left.

—D.O.

was that if the item had been more "difficult," and thus intended for "abler" students, the ambiguity in it would have been less objectionable. But all she was *really* saying was that she would have been more inclined to defend the item if fewer people had answered it correctly. (Or, to put it another way, she would have thought it less ambiguous if it had been more ambiguous.) This, of course, doesn't make any sense. Cruise had forgotten the real meaning of delta and fallen victim to her own circular logic. ETS's test developers cloak their work in scientific hocus-pocus and end up deceiving not only us but themselves.

Curley didn't share Cruise's peculiar concern. "Think we can defend," he wrote. "Words are in dictionary, they have modern usage, and we test more specialized *science* vocab than this. Aren't we willing to say that knowledge of these terms is related to success in college?"

Actual minorities

Whether the SAT is culturally biased against minorities is a perennial concern at ETS. The company says it has proven statistically that the SAT is fair for all. Just to make sure, for the last few years it has used "an actual member of a minority" (as one ETS employee told me) to read every test before it is published. According to an ETS flier, "Each test is reviewed to ensure that the questions reflect the multicultural nature of our society and that appropriate, positive references are made to minorities and women. Each test item is reviewed to ensure that any word, phrase, or description that may be regarded as biased, sexist, or racist is removed."

But the actual "sensitivity review" process is much more cursory and superficial than this description implies. The minority reviewer, a company employee, simply counts the number of items that refer to each of five "population subgroups" and enters these numbers on a Test Sensitivity Review Report Form.

On the verbal SAT administered in May 1982, minority member Beverly Whittington found seven items that mentioned women, one that mentioned black Americans, two that mentioned Hispanic Americans, none that mentioned native Americans, four that mentioned Asian Americans (actually, she was stretching here; these particular Asian Americans were Shang Dynasty Chinese, 1766—1122 B.C.). Two items overlapped, so Whittington put a "12" in the box for Total Representational Items. She also commented "OK" on the exam's text specifications, "OK" on the subgroup reference items, and "OK" on item review. She made no other remarks. If she had found the word *nigger* in one of the questions, presumably she would have scratched it out. ETS made Whittington take a three-day training program in "test sensitivity" before permitting her to do all of this. When her report was finished, it was stamped E.T.S. CONFIDENTIAL and SECURE. Then it was filed and forgotten.

Noun schmoun

After its sensitivity review, every SAT is passed along to what the College Board (which hires ETS to write and administer its college admissions tests) describes as a committee of "prominent specialists in educational and psychological measurement." ETS and the board talk publicly about the SAT committee as though it were a sort of psychometric Supreme Court, sitting in thoughtful judgment on every question in the SAT. According to the official mythology, the SAT committee ensures the integrity of the test by subjecting it to rigorous, independent, expert scrutiny. But in fact committee members are largely undistinguished in the measurement field. They have no real power, and ETS generally ignores their suggestions.

"They always hate to see my comments," says Margaret Fleming, one of the committee's ten members and a deputy superintendent in Cleveland's public school system. "Now, we have had some showdowns about it. Sometimes they change, but I find that item writers are very pompous about their work, and they don't like you to say anything. I am saying something, though, because I feel that maybe 40 people are responsible for writing items, let's say, for the verbal area, and why should 40 people govern by chance what thousands of youngsters' opportunities might be?"

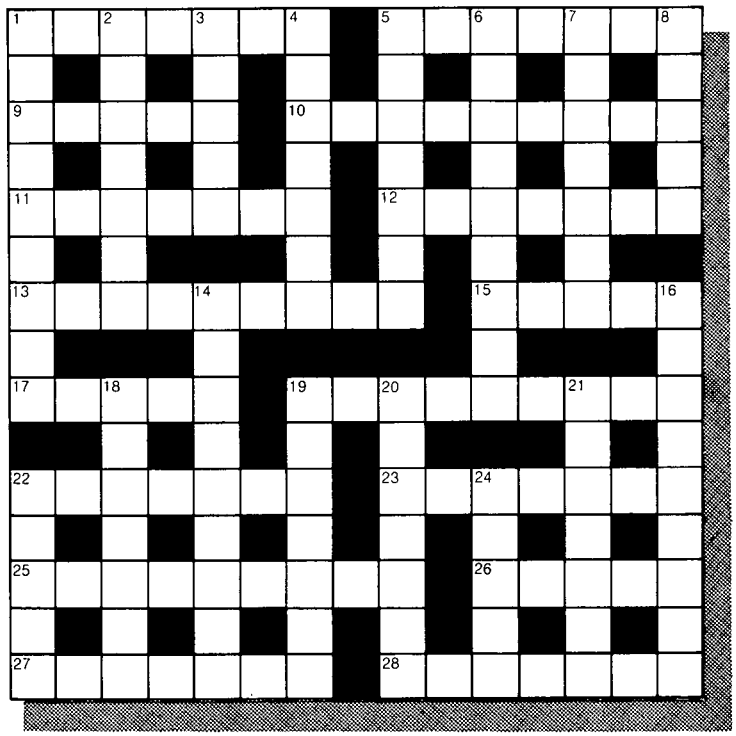
Far fewer than 40 people are involved in writing an SAT, but no matter. I asked her how ETS responded to her criticisms.

"They many times try to dismiss it," she said. "Sometimes they're very stubborn. Now, the one I got back recently was about a word that my dictionary said is a noun. Now I'm using *American Heritage Dictionary*, which I feel is common access across the country. I haven't got the Oxford English unabridged 30-volume thing. All I've got is what most people would have. And I said, the options here are verbs, and it appears that this

POLITICAL PUZZLE

by John Barclay

The numbers indicate the number of letters and words, e.g., (2, 3) means a two-letter word followed by a three-letter word. Groups of letters, e.g., USA, are treated as one word.



ACROSS

1. Study fast to ratify. (7)
5. Excellence seemed strange for Reagan appointee. (2, 5)
9. Steno transcribed shorthand. (5)
10. Roadside sign erroneously rocks putt. (5, 4)
11. Displayed a recent legislator. (7)
12. Shows up unseemly as paper. (7)
13. Political neutrality is mean thing. (3, 6)
15. Caper on the far side of the Atlantic. (5)
17. Clod puts four in Mrs. Dole's charge. (5)
19. Vessel set up to take Lett East. (9)
22. Crimes seen returning in this Red rumor. (7)
23. No encouragement from messy diapers. (7)
25. Another splintered rift freed. (9)
26. After 50, the bad would be good. (5)
27. New dew role brought down. (7)
28. Major political problem if cited incorrectly. (7)

DOWN

1. Agreed to putting off decent son. (9)
2. Tainter mixed fertilizer component. (7)
3. Stein rediscovered map feature. (5)
4. Thumb up, rode around, and rode around again. (7)
5. Teach in a new way, at deuce. (7)
6. Mediator's assignment from Thackeray? (4, 5)
7. Derive from former publication? (7)
8. Post office sex scandal for Montreal group. (5)
14. Rain around that place, not at the front. (2, 3, 4)
16. Vehicle disturbing the clover. (9)
18. With no eyes, fiery view can produce not much. (4, 3)
19. Kind of tuppence? (2, 5)
20. Duet aid conformed and confirmed. (7)
21. Intercourse if craft is accommodated. (7)
22. For decoration put five hundred in the repast. (5)
24. Assign people to bring Minnesota up before failing grade. (5)

Answers to last month's puzzle: