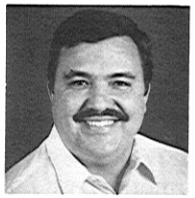
MEASURING DIFFICULTY and DIAGNOSTICITY in the AHSME MULTIPLE CHOICE EXAM

Terry Harman, Betty Reiter, Harold Reiter & Nancy Schoeps



Terry Hartman received a BS in Math with a Computer Science option in 1980 from the University of North Carolina-Charlotte. Since then he has been employed at UNC-Charlotte, currently as a VMS Systems Programmer. His interests are PCs and distributed processing.

Betty Reiter is Instructer of Mathematics at Winthrop College in Rock Hill, South Carolina. She has taught mathematics and computer science at university level. She is state coordinator of the AJHSME and AHSME with her husband Harold.

Harold Reiter is Associate Professor of Mathematics at the University of North Carolina at Charlotte. He is active in state and national mathematics contest work at the junior high, high school and college level. He is chairman of the American High School Math Exam committee.

Nancy Schoeps is in the Department of Mathematics, University of North Carolina at Charlotte. She is coordinator of the statistical consulting service of the University called OSAM, the Office of Statistics and Applied Mathematics. She gained a Ph D in 1979, from Syracuse University in New York, USA, in Psychology (Psychological Measurement and Statistics).

Introduction

Multiple choice exams which include a correction for guessing, (i.e. a penalty for wrong answers, or equivalently, a reward for questions left unanswered) have become increasingly popular among math contest administrators, because it is believed that this scoring system provides more information about students' performance than the simple right-wrong, 1-0, scoring (Pollard & Clark,1989). The exam consists of 30 multiple choice items with 5 choices for each question. Points for answers are awarded as follows:–

Right (or Correct)	(R)	5 points
Blank (or Omit)	(B)	2 points
Wrong	(W)	0 points.

It is intended that each student be aware of the scoring system and be advised to use that information to attain the maximum possible score. Three major characteristics of the test were analyzed:–

(a) difficulty of individual questions,

(b) diagnosticity of each question, and

(c) the overall reliability of the exam.

Each of these is discussed, in turn, below.

A. Difficulty of Individual Questions

To investigate difficulty, three related measures were used. They were:

1. C, the traditional measure, that is, the proportion of correct answers among all examinees;

2. D, the number of persons answering the question correctly divided by the number of persons who attempted the question, and;

3. E, which gives 5 points for each correct answer plus 2 points for each blank (to reflect the bonus for not guessing) and then divides by the total number of participants.

Small values of any of these measures mean that a question is "hard". Each of the three measures of difficulty, C, D, and E is a special case

of Z. Below are the formulas for Z, C, D, and E, where i is the item number. Remember that R=right, W=wrong, and B=blank.

$$Z_i = \frac{aR_i + bB_i}{R_i + cB_i + W_i}$$

$$C_i = \frac{R_i}{R_i + B_i + W_i}$$

$$D_i = \frac{R_i}{R_i + W_i}$$

$$E_i = \frac{5R_i + 2B_i}{R_i + B_i + W_i}$$

Each of the measures gives different information about the questions and/or the examinees. Let's look at an example of E. Suppose that 0.25 of the participants answered problem i correctly (R), 0.60 left it blank (B), and 0.15 answered it wrong (W). Then the average score on that question would be 2.45; compare that to a value of 0 if everyone answered it wrong, 2 if everyone left it blank, and 5 if everyone answered it correctly. Table 2 reports the values of C, D, and E for each question.

Although the three measures are variations of the formula for Z above, they are not perfectly correlated except in the case where everyone is instructed to answer all questions and there is no penalty for guessing. Since that is not the case in this exam, the three difficulty measures are not perfectly correlated. The correlations among C, D, and E are reported in Table 3. Because some of the measures were not normally distributed, the correlation coefficient calculated was a Spearman rank-order coefficient, rather than a Pearson.

B. Diagnosticity

Diagnosticity is taken here to be the extent to which an individual question predicts a high score on this whole exam. Five measures of diagnosticity used were PB, DIS, CC, A, and M. These are described here, discussed below, and reported in Table 6. For all five measures, a higher score indicates greater diagnosticity.

- PB is a standard method, called the point biserial index and is a measure that correlates each person's score on the item with

Table 2. Measures of Difficulty*

	C		D		E	
1	0.823	(29)	0.841	(29)	4.157	(29)
2	0.312	(24)	0.331	(19)	1.674	(6)
3	0.572	(28)	0.600	(26)	2.954	(26)
4	0.867	(30)	0.930	(30)	4.471	(30)
5	0.546	(27)	0.813	(28)	3.387	(28)
6	0.190	(20)	0.292	(16)	1.649	(4)
7	0.431	(26)	0.763	(27)	3.024	(27)
8	0.250	(22)	0.511	(24)	2.272	(24)
9	0.178	(19)	0.243	(11)	1.424	(2)
10	0.051	(11)	0.311	(17)	1.928	(19)
11	0.068	(14)	0.287	(15)	1.866	(17)
12	0.065	(12)	0.237	(10)	1.775	(11)
13	0.294	(23)	0.597	(25)	2.484	(25)
14	0.028	(6)	0.183	(9)	1.837	(15)
15	0.378	(25)	0.442	(23)	2.178	(23)
16	0.210	(21)	0.409	(22)	2.023	(22)
17	0.122	(17)	0.271	(13)	1.709	(8)
18	0.095	(16)	0.408	(21)	2.009	(21)
19	0.030	(8)	0.059	(1)	1.126	(1)
20	0.024	(3)	0.091	(3)	1.586	(3)
21	0.044	(10)	0.162	(8)	1.678	(7)
22	0.066	(13)	0.280	(14)	1.860	(16)
23	0.037	(9)	0.248	(12)	1.887	(18)
24	0.080	(15)	0.379	(20)	1.977	(20)
25	0.025	(4)	0.160	(6)	1.812	(13)
26	0.029	(7)	0.135	(5)	1.715	(9)
27	0.022	(2)	0.098	(4)	1.659	(5)
28	0.178	(18)	0.317	(18)	1.766	(10)
29	0.010	(1)	0.084	(2)	1.805	(12)
30	0.025	(5)	0.161	(7)	1.813	(14)

*Numbers in ( ) are the rankings of questions from easiest (30) to most difficult (1).

Table 3. Correlation Matrix of C, D and E
Correlation Analysis

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
C	30	0.201	0.235	6.049	0.010	0.867
D	30	0.354	0.239	0.640	0.059	0.929
E	30	2.116	0.760	63.506	1.126	4.471

Spearman Correlation Coefficients

	C	D	E
C	1.000	0.929	0.626
D	0.929	1.000	0.826
E	0.626	0.826	1.000

his/her score on the whole test. The PB was calculated for the entire group and for subsets of the group scoring over 60, 70, 80, 90, and 100 points.

$$PB_i = \frac{\overline{Y_i} - \overline{Y}}{s_y}\sqrt{\frac{p_i}{1-p_i}}$$

where
$\overline{Y_i}$ = the mean score of persons who answered the question right,
$\overline{Y}$ = the mean of all the scores,
s_y = the standard deviation of the scores, and
p_i = the proportion of people who answereed the question right.

- DIS is another standard measure, called the discrimination index. It subtracts the mean of the low scoring group from the mean of the high-scoring group and then divides that difference by 5. The high-scoring group usually represents from the top 10% to 50% of the participants, and the low-scoring group is the same percent of participants at the lower end. For tests where each item is scored only 0 or 1, this formula reduces to the proportion of the high-scoring group who answered this item correctly minus the proportion who answered correctly in the low-scoring group.

Choosing the percent of participants to include in the high and low scoring groups depends on the shape of the distribution. For normal distributions, the optimal percent is 27% in each group. For this group, 25% was used, a number suggested by the rule of thumb in *Introduction to Measurement Theory* by Allen and Yen.

- M is the mean score of all students whose answer to problem i was correct. It is to be expected that the mean score of students who answer a question correctly will be higher than the mean score of those who answer it incorrectly or those who leave it blank. Also, the higher mean scores should come from questions that are answered correctly by fewer people, those people being those who are in the high scoring group.

- A is the number of students who scored over 100 and got problem i correct divided by the number of students who answered problem i correctly.

- CC is the contingency coefficient which is based on the chi-square statistic for the question dividing the group into those scoring over 100 vs those scoring under 100 and then looking at the proportion of each group which scored 0, 2, or 5. The value of the CC ranges from 0 to 1 and is interpreted as a correlation coefficient. An example for one question follows:– Question 2

	Wrong(0)	Blank(2)	Right(5)
< 100	3979	361	1879
> 100	2	0 90	

Chi-square is 193.079 with 2 degrees of freedom. The contingency coefficient, CC, is 0.172, calculated by the following formula.

$$CC_i = \sqrt{\frac{\chi^2}{N+\chi^2}}$$

where N is the number of persons taking the test. The contingency coefficient is a way of looking at whether answering the question correctly is a good predictor of high test score overall. This question is an easy question; however, although almost everyone who scored over 100 answered it correctly, a large number of people

who scored below 100 also answered it correctly. Of course a very large number of persons answered it wrong and some people left it blank.

The five diagnosticity measures are compared by two approaches: first by pairwise correlations reported in Table 7; next, by lists of questions which showed up on all 5, or 4, 3, or 2 of the top ten lists for the five measures.

C. Reliability

The reliability of the test was assessed by Cronbach's coefficient alpha, a standard method for finding the reliability. Reliability as presented here is the traditional test characteristic that would be reported in any publication manual for a standardized test. The reliability of a test in classical test theory is a correlation between the observed score on the test and the true score. Cronbach's alpha is a lower bound estimate of this reliability, and can take on values between -1 and 1, although negative values are rare and undesirable. Reliability was calculated for the 0-2-5 scoring as well as for the 0,1 scoring, to compare the two systems.

Results and Discussion

What can be made of this sea of numbers? This section summarizes our pondering, and points to interesting conclusions about the topics introduced above: (a) difficulty, (b) diagnosticity, and (3) reliability.

A. Difficulty

Table 2 has information on the difficulty of the 30 questions. The intent of the test constructors is to make the questions of increasing difficulty. Items among the first 5 usually involve some arithmetic, order of operations, and technical meaning of a term (for example, what is an equiangular quadrilateral).

Items among the next fifteen require some ingenuity, but are accessible by most high school students who have studied algebra and geometry. The final ten problems on the test are generally considered hard by students. They are intended to differentiate among the top 1% of students. Each of the three different criteria gave a different ranking of questions from least to most difficult, none of them starting with question 1 and going

to question 30 in that order. Table 4 summarizes the items which are among the ten hardest as determined by all three measures, those which belong to two of the hardest ten lists, and those which belong to only one of the hardest ten list. The three statistics agree very well on problems among the first ten of the exam. Since statistics C and D are identical on problems which are answered by all participants, the values of C and D are close on problems for which the number of blanks (B) is small. Except for problem 15, which was left blank by only 14.3% of participants, the eight problems left blank least often are all among the first ten. Statistic E is just five times statistic C on problems no students leave unanswered; therefore, E and C are highly correlated on the early problems. For example, problem 3 was ranked 3rd easiest (28th hardest) by measure A and 5th easiest by the other two measures. Thus, it is not surprising to find that problems 3, 5, 7, 8, 13, and 15 did not differ across measures by more that two positions. At the other end of the test, the statistics differ greatly. The reasoning is similar to that above. Among the problems numbered 21 to 30, the one attempted most often was number 28, which had 56.2% attempt it (43.8% blanks and 12.4% wrong). The next most attempted problem in this group was problem 21, which was attempted by only 27% of the participants. The percent of correct answers among the last ten questions varies from 1% (problem 29) to 18% (problem 28). This huge relative difference among these final ten accounts for large differences among the three statistics, C, D, and E.

We believe that the statistic C is the most accurate and that E is the least accurate in measuring difficulty. We think that C is the best measure because a question which is worked correctly by a very small group of students (and missed by an even smaller group) but which, because it is highly complex or technical, is skipped by large numbers of students, should qualify as hard. Such questions contribute very little to the success of those students in the range 90 to 110, but would have a high D value (i.e., would be considered easy by measure D). Considering statistic E, if we could be sure that all students were well acquainted with the scoring system, with its highly negative effect for guessing, we could then assume that each student's decision to provide an answer in an attempt to earn five points was considered against the sure two points which could have been earned of leaving the problem unanswered. This

Table 4. Ten Hardest Problems by C, D, and E*

	C		D		E	
19	0.030	(8)	0.059	(1)	1.126	(1)
20	0.024	(3)	0.091	(3)	1.586	(3)
21	0.044	(10)	0.162	(8)	1.678	(7)
26	0.029	(7)	0.135	(5)	1.715	(9)
27	0.022	(2)	0.098	(4)	1.659	(5)
The problems above were rated hardest by all three measures						
14	0.028	(6)	0.183	(9)	1.837	(15)
25	0.025	(4)	0.160	(6)	1.812	(13)
29	0.010	(1)	0.084	(2)	1.805	(12)
30	0.025	(5)	0.161	(7)	1.813	(14)
The problems 14, 25, 29, and 30 were rated hardest by two measures						
2	0.312	(24)	0.331	(19)	1.674	(6)
6	0.190	(20)	0.292	(16)	1.649	(4)
9	0.178	(19)	0.243	(11)	1.424	(2)
12	0.065	(12)	0.237	(10)	1.775	(11)
17	0.122	(17)	0.271	(13)	1.709	(8)
23	0.037	(9)	0.248	(12)	1.887	(18)
28	0.178	(18)	0.317	(18)	1.766	(10)

*Numbers in ( ) are the rankings from easiest (30) to hardest (1).

assumption would make E a more credible measure. The data show that many persons answered all 30 questions. Most of them were low scorers leading us to the conclusion that they were guessing randomly, i.e., had not eliminated some of the distractors before choosing to guess. Answering patterns were highly variable, with some persons attempting harder problems at the end while skipping over easier problems. In Table 2 it can be seen from the values of E that many people attempted questions that they could not answer correctly. This is clear using statistic E. E would have a value of 2 if every participant left the question blank and a value of 1 if everyone guessed. The value of 1 comes from the fact that if everyone guesses, 1/5 of the guesses will be correct. Problem 19 has an E value just a little higher than 1, namely 1.126. Explanations for such a low value can be suggested, e.g., the diagram accompanying the problem may have led students to think they understood more than they did. The results of Tables 2 and 4, containing the correlations among the measures of difficulty, and the hypotheses about why such low values of E as we saw in question 19 might come up led to the definition of two more statistics, N and P. The rationale for these statistics was as follows. The correlation of D with C and E was higher than that of C with E. It seemed that C differed from D and E in that the problems for which the C value is small all look difficult; some of these may use a word or phrase with which students may be unfamiliar. We decided to compare questions using Nand P. Table 5 contains these results for each problem.

N_i = the number who answered question i Wrong divided by the number who answered it Not Right.

P_i = the number who attempted question i divided by the total number of participants.

If the percentage of blanks is large compared to the percentage of wrong answers, then the values of N and P will be small, which we expect to happen on problems which look hard. These two statistics, N and P, enable us to classify the hardest problems into two categories: the hard problems which look easy 19, 28, 9, 6, and 2 the problems which look hard 27, 14, 29, 30, 26, and 25. Problems in the "looks easy" category are those with high N and P values. These problems are of several types. Two are arithmetic (2,6), one is algebra (9), and the other two (19,28) are respectively geometry and combinatorics. Problem 2 is especially

Table 5. Values of N and P by problem

	P		N	
1	0.978	(30)	0.874	(28)
2	0.943	(28)	0.917	(30)
3	0.953	(29)	0.891	(29)
4	0.933	(27)	0.496	(23)
5	0.671	(24)	0.276	(17)
6	0.651	(23)	0.569	(25)
7	0.564	(22)	0.235	(14)
8	0.490	(17)	0.319	(19)
9	0.733	(25)	0.676	(26)
10	0.163	(6)	0.118	(3)
11	0.237	(12)	0.181	(10)
12	0.276	(15)	0.226	(13)
13	0.492	(18)	0.281	(18)
14	0.151	(3)	0.126	(4)
15	0.856	(26)	0.768	(27)
16	0.514	(20)	0.384	(21)
17	0.450	(16)	0.373	(20)
18	0.233	(10)	0.153	(8)
19	0.513	(19)	0.498	(24)
20	0.268	(13)	0.249	(16)
21	0.270	(14)	0.237	(15)
22	0.234	(11)	0.180	(9)
23	0.148	(2)	0.116	(2)
24	0.211	(7)	0.143	(7)
25	0.156	(5)	0.135	(6)
26	0.215	(8)	0.192	(11)
27	0.225	(9)	0.208	(12)
28	0.561	(21)	0.466	(22)
29	0.123	(1)	0.114	(1)
30	0.156	(4)	0.134	(5)

interesting. It asks the student to evaluate the expression $|3 - \pi|$. Its N and P values were 0.943 and 0.917 respectively. As it deals with absolute value and irrationality, one might guess that most students who miss this problem would miss it because of a misunderstanding of absolute value. In fact, only 10% of the population selected distractor (c) which targets the understanding of the sign in absolute value, while another 8% selected alternative (d), which would be a more basic misunderstanding, that of order of operations,an attempt to "distribute absolute value". Instead most students (38%) missed the problem because they treated 3.14 as the exact value of π. This is the second problem in the test and should therefore be one of the easiest. Yet only 31% of the students got it right as compared to questions 3 and 4 for which the percent who answered correctly was 57 and 87 respectively. In the "looks hard" category we find one problem each in number theory (14) and geometry (29), and two problems each in algebra (25,27) and combinatorics (26,30). For these problems the N and P values are all less than 0.25.

B. Diagnosticity

The results of the diagnosticity analysis show that the five criteria are tapping different aspects of predicting success on the exam. Table 6 shows the values of each of the five statistics for each of the thirty questions. PB, DIS, CC, and A range between -1 and 1 where values closer to 1 indicate that the item does a good job predicting the total score on the exam. The information in Table 7 is the correlation among the 5 measures of diagnosticity. Although the five statistics are positively correlated, A is the least correlated with all of the others. In fact, the correlations among PB, CC, and M are all over 0.60. The correlations of DIS with PB and M is high, but the correlation with A and DIS is actually negative. The correlation of A with CC is high, but the corrlation of A with each of the others is low. Which of these statistics is the best predictor of success? Unfortunately, none does this very well by itself. The correlation of DIS with the others? DIS, the measure of discrimination, is higher for the easier questions and is positive for each question. As with the point biserial, the value of the discrimination is limited by the difficulty of the question, with very hard or very easy questions having a restricted value of the discrimination index. The D measure is particularly suspect here because more than 25% of the participants scored less than 60, a score, as noted earlier, would be

achieved by leaving all the questions blank.

The point biserial measure values were somewhat low for a standardized test; usually values from 0.3 to 0.7 are expected on a test for general use. However, this is such a difficult test that it is not surprising that the restriction of range would lead to lower correlations; it can be argued that this is not a test for general use. That makes it difficult to say how large the PB should be, especially since the interpretation of the size of the PB depends on the proportion of people who got it right. When comparing the groups of all students who had scores greater than 0, greater than 60, greater than 70, greater than 80, greater than 90, and greater than 100, the following pattern emerged. For the first ten questions the point biserial decreased as the group became increasingly selective. However, among the last ten problems, which are the most difficult, the PB increased as the group became more selective, and in nearly all cases this was a monotone increase. An exception was Problem 24 where the PB was between 0.4 and 0.5 for all groups except the group that scored 100 or higher, for which the PB was 0.324. The statistic M seemed to do the best job with the easier problems. In fact, using that criterion, seven of the "best" problems were among the first 15 on the test. Using the A statistic the better questions came from the middle 10, which were of much lower difficulty than the first 10. The contingency coefficient, CC, also rates the middle 10 problems among its best, although it does have 3 from the first 10 and 2 from the last ten. One important reason to study this topic is to better understand what distinguishes mathematically talented students from their peers, and to more accurately identify these students. It might be argued that an exam of the nature of the AHSME should test reasoning power and ingenuity, and it should avoid problems which require a high degree of technical understanding or formulas. However, our findings show that some problems involving technical knowledge are excellent predictors of success on the exam. Problem 24, which involves the rotation of a graph of a function is a case in point.

Since none of the statistics for diagnosticity predicted as well as we would like, we tried some other approaches. We speculated that the number of questions attempted would be a good predictor of success. In fact, we thought that a disproportionately large number of Honor Roll (score of 100 or more) students would attempt exactly 15 questions. A score of

Table 6. Measures of Diagnosticity*

	A	CC	PB	M	DIS
1	0.018(3)	0.056(3)	0.272(15)	65.6(12)	0.298(21)
2	0.046(10)	0.172(21)	0.561(28)	76.1(25)	0.666(29)
3	0.025(4)	0.097(12)	0.501(26)	70.2(17)	0.672(30)
4	0.016(2)	0.040(2)	0.275(16)	65.1(10)	0.211(18)
5	0.026(6)	0.108(13)	0.485(23)	69.1(14)	0.411(23)
6	0.075(22)	0.237(28)	0.564(29)	80.0(30)	0.485(25)
7	0.033(7)	0.133(14)	0.489(24)	70.0(16)	0.394(22)
8	0.058(14)	0.203(24)	0.623(30)	78.5(28)	0.545(27)
9	0.075(21)	0.228(27)	0.543(27)	79.4(29)	0.543(26)
10	0.103(24)	0.169(19)	0.168(11)	64.3(8)	0.064(3)
11	0.138(28)	0.268(30)	0.363(19)	77.6(26)	0.168(15)
12	0.104(25)	0.194(23)	0.222(14)	69.4(15)	0.138(13)
13	0.046(11)	0.169(18)	0.498(25)	73.6(23)	0.459(24)
14	0.149(30)	0.186(22)	0.166(10)	66.4(13)	0.084(6)
15	0.035(8)	0.134(15)	0.423(22)	72.2(21)	0.588(28)
16	0.065(16)	0.212(25)	0.411(21)	71.5(19)	0.289(20)
17	0.070(19)	0.171(20)	0.276(18)	71.4(18)	0.215(19)
18	0.068(18)	0.146(17)	0.187(12)	64.1(7)	0.091(8)
19	0.063(15)	0.080(8)	0.102(4)	63.0(6)	0.046(2)
20	0.065(17)	0.082(9)	0.161(8)	71.7(20)	0.124(12)
21	0.141(29)	0.220(26)	0.275(17)	75.8(24)	0.101(10)
22	0.036(9)	0.069(5)	0.061(1)	58.1(2)	0.077(4)
23	0.056(12)	0.074(7)	0.158(7)	64.7(9)	0.083(5)
24	0.127(27)	0.266(29)	0.400(20)	77.9(27)	0.174(16)
25	0.057(13)	0.057(4)	0.112(6)	60.9(4)	0.042(1)
26	0.071(20)	0.086(11)	0.107(5)	59.1(3)	0.090(7)
27	0.115(26)	0.134(16)	0.189(13)	72.6(22)	0.167(14)
28	0.026(5)	0.083(10)	0.162(9)	65.4(11)	0.205(17)
29	0.092(23)	0.072(6)	0.101(3)	61.5(5)	0.101(11)
30	0.006(1)	0.030(1)	0.062(2)	55.1(1)	0.094(9)

*Numbers in ( ) are the rankings from 1 least diagnostic to 30 most diagnostic.

100 is attained by correctly answering 14 questions in 15 attempts. If a student attempts 16 and gets 14 correct, that student scores less than 100. On the other hand a student who answers 14 questions must get every one right to score 100 and so has little to lose by guessing at one more. However, the number of students answering exactly 15 questions is not substantially different from the expected number of students in that group. To find the expected number of students attempting n questions, $0 \leq n \leq 30$, we first found the probability p_i that problem i is attempted. Then for each n element subset A of $\{1, 2, \ldots, 30\}$, we calculate the probability P_A that a randomly selected student attempts exactly the problems in A. Then the probability P_n that a random student attempts exactly n problems is:–

$$P_A = \prod_{i=1}^{30} x_i \quad \text{where } x_i = \begin{cases} p_i & \text{if } i \in A \\ (1 - p_i) & \text{if } i \notin A \end{cases}$$

$$\sum \{P_A : A \text{ is an } n \text{ element subset of } \{1, 2, ..., 30\}\}.$$

These calculations depend on the admittedly difficult to justify assumption that the event that problem i is attempted is independent of the event that problem j is attempted for $i \neq j$. Even when we restricted our analysis to those students with scores in the 90 to 109 range, we did not find a significant chi-squared value for the difference between the number of students who attempted 15 questions and the expected number. Also, the average score sorted by number of attempts fails to show that score and number of attempts is highly correlated either linearly or quadratically, although the quadratic model is a slightly better fit. For the linear model we get an R-square of 0.055 and for the quadratic 0.111.

C. Reliability

In calculations of the reliability of the test, Cronbach's alpha was 0.736 for the 0-2-5 scoring. Reliabilities of 0.90 to 1.00 are considered high, 0.80 to 0.90 moderate, and 0.70 to 0.80 low. Reliabilities lower than 0.70 are considered to be a problem on standardized tests. The AHSME alpha, therefore, was low. Part of the reason for the lower alpha might be that the high test difficulty restricted the range. When the alpha was computed using the 0-1 scoring the reliability was not very different, 0.717. It seems, therefore, that the correction for guessing does not appreciably increase the reliability of the AHSME.

Table 7. Correlation matrix of CC, PB, M, Dis and A
Correlation Analysis

Simple Statistics

Variable	N	Mean	Std Dev	Minimum	Maximum
A	30	0.067	0.039	0.006	0.149
CC	30	0.139	0.069	0.030	0.268
PB	30	0.297	0.17	0.061	0.623
M	30	69.010	6.732	55.10	80.00
DIS	30	0.254	0.200	0.040	0.670

Spearman Correlation Coefficients

	A	CC	PB	M	Dis
A	1.000	0.668	-0.037	0.327	-0.309
CC	0.668	1.000	0.637	0.793	0.353
PB	-0.037	0.637	1.000	0.832	0.878
M	0.327	0.793	0.832	1.000	0.695
DIS	-0.309	0.353	0.878	0.695	1.000

Table 8. Most Diagnostic Problems, Composite of All Five Measures

6	0.075(22)	0.237(28)	0.564(29)	80.0(30)	0.485(25)
9	0.075(21)	0.228(27)	0.543(27)	79.4(29)	0.543(26)
The problems above were ranked in the top ten by all five measures					
2	0.046(10)	0.172(21)	0.561(28)	76.1(25)	0.666(29)
8	0.058(14)	0.203(24)	0.623(30)	78.5(28)	0.545(27)
11	0.138(28)	0.268(30)	0.363(19)	77.6(26)	0.168(15)
13	0.046(11)	0.169(18)	0.498(25)	73.6(23)	0.459(24)
15	0.035(8)	0.134(15)	0.423(22)	72.2(21)	0.588(28)
21	0.141(29)	0.220(26)	0.275(17)	75.8(24)	0.101(10)
24	0.127(27)	0.266(29)	0.400(20)	77.9(27)	0.174(16)
The problems above were ranked in the top ten by three or four of the measures					

Summary

In the AHSME the measures of difficulty that we used gave fairly consistent results. The inclusion of the idea of problems that look hard as opposed to those that look easy but actually are hard was a new way to look at difficulty and gave us some additional information about the performance of students on the test. In investigating diagnosticity, none of the single statistics we used predicted success on the test as a whole very well. We did, however, find that certain problems were good at predicting score on the test, and those questions tended to be good by all five or several of the five criteria. See Table 8. We found that the number of problems a student attempts does not help in predicting the high scorers, although there is a bit of a pattern around 15 to 17 attempts, which gives those students a slightly higher average score. The reliability of the AHSME falls in the acceptable, although low, range, and is not significantly different for 0-2-5 scoring than for 0-1 scoring. Overall this look at ways of analyzing difficulty and diagnosticity gave an interesting view of students' behavior on this difficult competitive exam.

Table 1. Demographics

	Males	Females	Total
Asians	213	181	394
Blacks	14	226	372
Hispanics	35	12	47
Whites	2732	2421	5153
TOTALS	3126	2840	5966

Grade	Males	Females	Totals
8 or less	63	46	109
9	233	264	497
10	735	692	1427
11	941	886	1827
12	1211	996	2207
TOTALS	3183	2884	6067

AHSME problems

1. If for any three distinct numbers a, b, and c we define $\boxed{a;b;c}$ by

$$\boxed{a;b;c} = \frac{c+a}{c-b} \quad \text{then} \quad \boxed{1;\ -2;\ -3} =$$

(A) -2 (B) $-\frac{2}{5}$ (C) $-\frac{1}{4}$ (D) $\frac{2}{5}$ (E) 2

2. $|3-\pi| =$

(A) $\frac{1}{7}$ (B) .14 (C) $3-\pi$ (D) $3+\pi$ (E) $\pi-3$

3. $(4^{-1}-3^{-1})^{-1} =$

(A) -12 (B) -1 (C) $\frac{1}{12}$ (D) 1 (E) 12

4. Which of the following triangles cannot exist?

(A) An acute isosceles triangle (B) An isosceles right triangle
(C) An obtuse right triangle (D) A scalene right triangle
(E) A scalene obtuse triangle

6. If $x \geq 0$, $\sqrt{x\sqrt{x\sqrt{x}}} =$

(A) $x\sqrt{x}$ (B) $x\sqrt[4]{x}$ (C) $\sqrt[8]{x}$ (D) $\sqrt[8]{x^3}$ (E) $\sqrt[8]{x^7}$

7. If $x = \frac{a}{b}$, $a \neq b$ and $b \neq 0$, then $\frac{a+b}{a-b} =$

(A) $\frac{x}{x+1}$ (B) $\frac{x+1}{x-1}$ (C) 1 (D) $x - \frac{1}{x}$ (E) $x + \frac{1}{x}$

10. Point P is 9 units from the center of a circle of radius 15. How many different chords of the circle contain P and have integer lengths?

(A) 11 (B) 12 (C) 13 (D) 14 (E) 29

11. Jack and Jill run 10 kilometers. They start at the same point, run 5 kilometers up hill, and return to the starting point by the same route. Jack has a 10-minute head start and runs at the rate of 15 km/hr uphill and 20 km/hr downhill. Jill runs 16 km/hr uphill and 22 km/hr downhill. How far from the top of the hill are they when they pass going in opposite directions?

(A) $\frac{5}{4}$ km (B) $\frac{35}{27}$ km (C) $\frac{27}{20}$ km (D) $\frac{7}{3}$ km (E) $\frac{28}{9}$ km

14. The measures (in degrees) of the interior angles of a convex hexagon form an arithmetic sequence of positive integers. Let $m°$ be the measure of the largest interior angle of the hexagon. The largest possible value of $m°$ is

(A) 165° (B) 167° (C) 170° (D) 175° (E) 179°

15. If x is the cube of a positive integer and d is the number of positive integers which are divisors of x, then d could be

(A) 200 (B) 201 (C) 202 (D) 203 (E) 204

17. A positive integer N is a *palindrome* if the integer obtained by reversing the digits of N is equal to N. The year 1991 is the only year in the current century with the following two properties: (a) It is a palindrome. (b) It factors as a product of a 2-digit

palindrome and a 3-digit palindrome. How many years in the millenium between 1000 and 2000 (including the year 1991) have properties (a) and (b)?

(A) 1 (B) 2 (C) 3 (D) 4 (E) 5

18. If S is the set of points z in the complex plane such that $(3+4i)z$ is a *real* number, then S is a

(A) right triangle (B) circle (C) hyperbola (D) line (E) parabola

20. The sum of all real x such that

$$(2^x - 4)^3 + (4^x - 2)^3 = (4^x + 2^x - 6)^3 \quad \text{is}$$

(A) 3/2 (B) 2 (C) 5/2 (D) 3 (E) 7/2

21. If $f\left(\frac{x}{x-1}\right) = \frac{1}{x}$ for all $x \neq 0, 1$ and $0 < \theta < \frac{\pi}{2}$, then $f(\sec^2\theta) =$

(A) $\sin^2\theta$ (B) $\cos^2\theta$ (C) $\tan^2\theta$ (D) $\cot^2\theta$ (E) $\csc^2\theta$

24. The graph, G, of $y = \log_{10} x$ is rotated 90° counter-clockwise about the origin to a new graph G'. Which of the following is an equation for G'?

(A) $y = \log_{10}\left(\frac{x+90}{90}\right)$ (B) $y = log_x 10$

(C) $y = \frac{1}{x+1}$ (D) $y = 10^{-x}$ (E) $y = 10^x$

25. If $T_n = 1 + 2 + 3 + \ldots + n$ and $P_n = \frac{T_2}{T_2 - 1} \cdot \frac{T_3}{T_3 - 1} \cdot \ldots \cdot \frac{T_n}{T_n - 1}$ for $n = 2, 3, 4, \ldots$, then P_{1991} is closest to which of the following numbers?

(A) 2.0 (B) 2.3 (C) 2.6 (D) 2.9 (E) 3.2

26. An n-digit positive integer is cute if its n digits are an arrangement of the set {1, 2, . . . , n} and its first k digits form an integer that is divisible by k, for $k = 1, 2, \ldots, n$. For example, 321 is a cute

3-digit integer because 1 divides 3, 2 divides 32, and 3 divides 321. How many cute 6-digit integers are there?

(A) 0 (B) 1 (C) 2 (D) 3 (E) 4

27. If

$$x + \sqrt{x^2 - 1} + \frac{1}{x + \sqrt{x^2 - 1}} = 20$$

then

$$x^2 + \sqrt{x^4 - 1} + \frac{1}{x^2 + \sqrt{x^4 - 1}}$$

equals

(A) 5.05 (B) 20 (C) 51.005 (D) 61.25 (E) 400

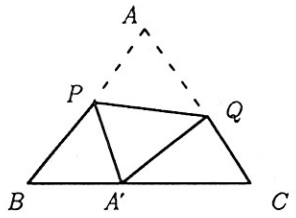
28. Initially an urn contains 100 black marbles and 100 white marbles. Repeatedly, three marbles are removed from the urn and replaced from a pile outside the urn as follows:–

Marbles Removed	Replaced With
3 black	1 black
2 black, 1 white	1 black, 1 white
1 black, 2 white	2 white
3 white	1 black, 1 white

Which of the following sets of marbles could be the contents of the urn after repeated applications of this proceedure?

(A) 2 black marbles (B) 2 white marbles (C) 1 black marble
(D) 1 black and 1 white marble (E) 1 white marble

29. An equilateral triangle ABC has been creased and folded so that vertex A now rests at A' on $\overline{BC}$ as shown. If $BA' = 1$ and $A'C = 2$, the the length of crease $\overline{PQ}$ is



(A) $\frac{8}{5}$ (B) $\frac{7}{20}\sqrt{21}$ (C) $\frac{1+\sqrt{5}}{2}$ (D) $\frac{13}{8}$ (E) $\sqrt{3}$

30. For any set S, let $|S|$ denote the number of elements of S, and let $n(S)$ be the number of subsets of S, including the empty set and the set S itself. If A, B, and C are sets for which

$$n(A) + n(B) + n(C) = n(A \cup B \cup C) \text{ and } |A| = |B| = 100,$$

then what is the minimum possible value of $|A \cap B \cap C|$?

(A) 96 (B) 97 (C) 98 (D) 99 (E) 100

References

Allen MJ & Yen WM, (1979), *Introduction to Measurement Theory*, Wadsworth.

Cocker L & Algina J, (1986), *Introduction to Classical and Modern Test Theory*, Holt, Reinhart and Winston.

Pollard GH & Clark DI, (1989), An Optimum Scoring System for Multiple Choice Competitions: an Analysis of Candidates Responses under Two Different Methods of Scoring, *Mathematics Competitions*, Vol 2 No 2, 33–36.

Nancy Schoeps,
Terry Harman,
Harold Reiter,
University of North Carolina
at Charlotte,
Charlotte, NC 28223,
USA.

Betty Reiter,
Winthrop University,
Rock Hill, SC, 29733,
USA.

* * *