ED 216 049 TM 820 242

AUTHOR Lutkus, Anthony D.; Laskaris, George
TITLE Item Analysis of Publisher-Supplied Test Questions in Introductory Psychology.
PUB DATE 25 Aug 81
NOTE 26p.; Paper presented at the Annual Meeting of the American Psychological Association (89th, August 25, 1981).

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Difficulty Level; Higher Education; *Item Analysis; *Item Banks; Multiple Choice Tests; Psychology; *Test Construction; Testing Problems; Test Items; *Test Selection

ABSTRACT

Analyses of student responses to Introductory Psychology test questions were discussed. The publisher supplied a two thousand item test bank on computer tape. Instructors selected questions for fifteen item tests. The test questions were labeled by the publisher as factual or conceptual. The semester course used a mastery learning format in which repeat testing was conducted using alternate test forms. Standard item analysis included the percentage of students passing and correlation coefficients. The second analysis was based on a technique used in designing the New Jersey College Basic Skills Placement Test. Each question was described as a function of the probability of a correct response for students at five ability levels as defined by total test scores. A difficulty curve for each item resulted which allowed the instructor to see if the question discriminated equally well among students at each ability level or whether the item presents problems only to students in a particular ability range. Test items with identical correlations with total test scores often yielded different difficulty curves. Difficulty curves with several different slope types were analyzed as a function of the syntactic "frame" of the question. (Author/DWH)

Item Analysis of Publisher-Supplied Test Questions In Introductory Psychology

Anthony D. Lutkus
and

George Laskaris

RUTGERS UNIVERSITY
Newark, N. J.

Paper Read at the 89th Annual Convention of the American Psychological Association, August 25, 1981.

Item Analysis of Publisher-Supplied Test Questions In Introductory Psychology

Anthony D. Lutkus
and
George Laskaris

Rutgers University - Newark

In 1932 Carl Brigham, father of the College Board's Scholastic Aptitude Test, published a small red volume titled "A Study of Error." Reviewers with almost fifty years hindsight (Donlon, 1979; Findley, 1981) point out to us that the hard nosed, data oriented Brigham was actually following in the "process" tradition of Binet when he suggested that one can (and indeed should) learn something about the workings of the mind by considering the patterns of wrong answers to multiple-choice test questions. Brigham actually developed diagrams much like a computer programmer's flow charts to represent possible thought patterns in the responses to his multiple-choice items. Today we would call him a "cognitive" psychologist. Brigham referred to his work simply as "digging" and invited others to do the same. It is in this spirit that we present the results of our "digging" into a 945 item test bank of multiple-choice test questions for Introductory Psychology.

Virtually every commercially successfully Introductory Psychology text provides the instructor a test file of multiple-choice questions. For the instructor with a large "intro" class, these test files are at once a blessing and a bane. Every publisher lauds the qualifications of his test maker but few provide data on expected difficulty in quantitative terms. Thus, the instructor who selects his questions from such a file must make "seat of the pants" guesses in editing-out suspect items. Less fortunate is the instructor employing a mastery learning or test-retest system.* He

[1] This research was sponsored, in part, by a grant from the Rutgers University Council on Instructional Development.

or she has no choice but to use almost all the supplied items in order to generate alternate test forms. The writers were in the latter situation but the use of a computer-managed mastery learning format has allowed us some hindsight into relative item difficulty.

Subjects

One hundred and seventy-five students in two one-semester introductory psychology classes provided the data by answering varying proportions of the 945-question test bank. Students self-selected the psychology course but did not know in advance that it would have a mastery learning format. Fifty-seven % were men and 43% were women. Students from this urban commuter campus at Rutgers-Newark have sat scores slightly above the national average for four-year colleges.

Procedure

The course required that students attempt the tests (15 items each) on at least 14 of 20 chapters in the text. Three different 15-item forms were available in random order for each chapter. Students who did not meet the criteria for "mastering" a chapter (13 of 15 correct) thus had two additional attempts available on the same material. The number of students responding to each question in the bank varied from 114 to as few as 33 on some alternate forms of optional chapters.

Students tested on a self-paced schedule. They recorded their responses to the four-choice test questions on a standard type scan form. The forms were optically scanned and the results processed by an on-line computer program as the student waited. Within a minute the student could view the results on a CRT computer terminal. The screen displayed the item numbers of all wrong answers, the correct answer and the page numbers in the textbook where both the correct answer and the distractors could be found. Students were allowed to keep their test forms while they viewed their results. All the error data were saved by the program for the present analysis.

The text used in the course was the second edition of Hall, Lindzey, and Thompson's Psychology, published by Worth in 1978. The 2000 item test bank was supplied by the publisher on a computer tape and questions for the 15-item tests were selected by the instructors and printed for student use. The publisher described the following features of the test bank:

1) Items were labeled as being either "factual" or "conceptual."
2) The "conceptual" items emphasize an understanding of the textbook material as it applies to situations and examples not given in the text. Many of the conceptual questions present an example and require the student to pick a term that best fits the example. Others present a term and require the student to select an example that best illustrates the term. Still other conceptual items require the student to make predictions based on information presented in the text, to integrate ideas, to solve analogies, or to apply the learned information to real or hypothetical situations not specifically mentioned in the textbook.
3) The factual questions require recognition of specific facts, definitions, or infonmation presented in the text.

Results

Two different modes of item analysis were used. The first or "standard" item analysis can be reviewed quickly. Figure 1 shows a frequency histogram of the "percent passing" for the 945 items. The mean percent passing for all items was 67.62%--very close to the traditional "ideal" average of a 70% "C" grade for a class. As can be seen in Figure 1, there were very few "poor" questions, only 11.6% of all the items had passing rates below 50%. The modal category for the test bank was 70-79% correct.

Since all the item responses were saved in a computer file, we were able to complete several analyses related to the publisher's stated characteristics of the test item bank. Analysis of variance revealed, interestingly, that there was no difference in overall difficulty between the questions called "conceptual" and those labeled "factual." It would probably be of little value for an instructor to spend time mixing, matching or proportioning his tests on the basis of these labels--a strategy that one might have been tempted to use if one expected "conceptuals" to be more difficult or thought provoking than "factual" types.

There was also no difference in mean item difficulty by chapter. This is an important point since it suggests that the question-analysis to follow here can be considered content-independent. The chapter mean percent passing ranged from 74.28% to 64.01% for the nineteen chapters. Ranking the first five chapters yields the following order with easiest first: (1) Sex, (2) Memory, (3) Language, (4) Behavior Disorders and (5) Evolution and Heredity. The chapter on human sexuality was, incidently, not required reading.

The second analysis involves a technique recently used to edit items for the New Jersey State College Basic Skills Placement Test. Test questions are usually charact rized by a single summary statistic, typically "percent passing" and/or the point biserial correlation coefficient. These single descriptors do not give the test maker information about the question's discriminating power across the range of students. In other words, the fact that a question was answered correctly sixty percent of the time; could mean that equally 60% of A, B, C and D level students passed it or that some combination of the A and B grade students passed and all the C and D students failed. One way to achieve more ~~detailed~~ information about what a test question does is to plot item/test regression curves. Figure 2 illustrates such a curve.

The ordinate shows the percentage of students who passed the item. The abscissa is divided into five categories based on the student's total test score on the form in which the test item appeared. Each curve represents the results for one question. If the "A" students, (as defined by the internal criterion of their total test score) have the highest probability of passing the item and the "B" students the next highest probability and so on, down to the poorest students, we would expect the linear ascending function shown in Figure 2. In fact Figure 2 is what the "average" question does look like for this test bank. All the item data were summated to generate Figure 2. The bands around each point indicate the standard deviations for the 945 items. Approximately one-third of the items plot within one standard deviation of this graph.

An easy question, or one with a "ceiling" effect is plotted in Figure 3. A, B, C and sometimes even D graded students are highly successful on this type of question. It does not discriminate well across the range of students.

An unfairly difficult question would have the slope type shown in Figure 4. We termed this a "flat-low" slope. This kind of question also is a poor discriminator in that everyone misses it--regardless of ability.

The fourth slope type is shown in Figure 5. Here the "A" and "B" students have reasonable probabilities of success--at least comparable to the linear slope from Figure 2--but the "C" and "D" students are at roughly the chance level. It is interesting to note that one could not predict whether a question was a "high end discriminator" or a "linear" type by knowing the single statistic of average percent passing.

We discovered each of these four curve types in the question bank by having the computer actually plot about a third of the items. Other types of slopes are possible; for example, a "medium-level discriminator"--given a different type of test. A placement test or minimum skills test where the questions can be hierarchically ordered (as in algebra) would tend to have medium level discriminators (Dass and Pine, 1981).

Once we had sampled the types of slopes in the item bank we wrote an algorithm to describe each curve mathematically. The computer was then used as a pattern recognition device to loop through the data for all the items and identify those questions which resembled each type within the bounds of .75 of a standard deviation at each category. Approximately 40% of the questions "fit" into one of these four slope types. This capture ratio could be improved if wider confidence limits are chosen but then the slope types tend to loose definition and overlap.

Table 1 presents a summary of the descriptive data for each of the slope types. Notice that the point biserial correlation coefficients range widely within each slope type and consequently would not give the tester a clue as to what kind of slope a given question might generate. The point biserial correlation coefficient is frequently used as a measure of the appropriateness of a test item relative to the total test. The Educational Testing Service (Hecht and Swineford, 1981) operates under the convention that moderate levels of mean biserial correlations (between .40 and .55) are "good" and low levels (less than .25 or .30) are suspect. Our data indicate that either level may yield a useful item, if you know its slope type.

In Table 1 it can be seen that knowing the percent passing rate can be a useful indicator of slope type but only if the rate is very high (ceiling type slope) or very low. Consequently we asked the question whether slope types on the item/test regression curves can be predicted by any other variables. Since the publisher's question writers deliberately used several question "frames" across all the items in this test bank we decided to code the questions according to their "form" (not content) and look for proportional patterns within slope types.

The question forms or frames we coded were:

Example into Term - The student is given a hypothetical example and required to choose the term that best fits the example.

An investigator measures the speed at which a frog catches a fly. The researcher is demonstrating

1. Repeatability.
2. Quantification.
3. Subjectivity.
4. Communication.

Two Blanks - A question with two fill-ins; the pair to be selected as a unit from four choices.

Kinsey reported differences among women's sexual behaviors. In general, college women engaged in more _____ and less _____ than less-educated women.

1. Premarital intercourse; homosexual behavior
2. Homosexual behavior; masturbation
3. Petting; masturbation
4. Masturbation; premarital intercourse

Definitions - The student has to recognize a definition either in the question body or from among the multiple choices.

Psychoanalysis is

1. A method of studying behavior.
2. A theory of behavior.
3. Both a method of studying behavior and a theory of behavior.
4. A survey method.

Recognizing facts or findings - The student must find or recognize a fact or result of a study.

The portion of the central nervous system that has been shown to play a role in impotence and certain fetishes is the

1. Septum.
2. Frontal Lobe.
3. Hypothalamus.
4. Temporal Lobe.

Integrations of Ideas - A question involving several conceptual steps, such as combining knowledge of several definitions or findings and relating them to material not in the text.

A normal adult requests a "split-brain" operation in order to increase her ability to perform two tasks simultaneously. A major argument against granting her request is that after such an operation

1. Her walking, running, and balance would be impaired.
2. She would be less intelligent.
3. Her speech would be disrupted.
4. The processes in each of the two halves of her brain would not be coordinated.

Prediction - The student is required to forsee results of a stated manipulation or to chose alternatives.

Jeff has a genetic predisposition toward TB. The likelihood that he will contract the disease is

1. Low, but only if he is careful.
2. High, because of his susceptibility.
3. High, if he is in contact with his relatives.
4. Low, because of improved living conditions.

Not true of - This category includes questions where the student is called upon to do an exclusion process. Questions here include sentences such as "all but which of" or "which is not true of...".

Kinsey's conclusions regarding differences between men and women included all but which of the following?

1. Men react more to erotic stories.
2. Men talk more about sex:
3. Men prefer more diverse forms of bodily stimulation.
4. Men engage in more sexual fantasies.

The relationship between these question frames and three of the item slope types is shown in Figures 6, 7, and 8. The percentage of each question frame type found within each slope category is plotted on the vertical axis. Note that both the linear type slope and the ceiling type contain large proportions of "recognizing facts" and "example-into-terms." The pattern for a "high-end discriminator" slope is different. The relative proportion of "example into term" frame types remains large but the proportion of factual recognitions decreases. This difference in patterns fits with the general notion that if a question discriminates only among the better students it probably involves some additional thought processes or mental steps beyond recognition of facts. This seems to be supported mainly by the decrease in proportion of factual recognitions but, unfortunately, not by any substantial increase in question frames that might require more complex (2 blank, integration, or prediction) thinking.

Conclusion

Based on our analyses we would recommend that test item publishers refrain from labeling their questions by unvalidated categories like factual vs conceptual. Second, we would like to see publishers furnish data on expected percent passing for each question. Better still would be information on expected slope type generated by each question. Since many publishers claim to pre-test their questions, this recommendation may not be as costly as it at first seems. Armed with a knowledge of the expected slope types an instructor could construct tests that would contain a known mix of average (linear) questions, as well as high-end-discriminators that would challenge the better students. Questions with ceiling type slopes could be judiciously sprinkled in as the "gifts" they are.

For the instructor faced with selecting questions from a test bank without such supplementary statistics we cannot yet offer a definitive method for moving from question frame to predicting slope type since all frame types can be found within each kind of slope. The important variable is the proportion of frame types that make up the test. We would suggest that questions involving the simple recognition of facts or findings should make up no more than 35 to 50% of a college level test. Questions involving going from examples into terms, or manipulating information in a way not found in the text are more likely to yield "high-end discriminators" that will challenge the better student.

REFERENCES

Dass, J. and Pine, C.; Interpreting Mathematics Scores on the New Jersey College, Basic Skills Placement Test, Princeton, N.J.: New Jersey Department of Higher Education, Basic Skills Council and the Educational Testing Service, 1981.

Diederich, Paul B., Short-Cut Statistics For Teacher-Made Tests. Educational Testing Service Monograph, 1973.

Donlon, Thomas F., Brigham's Book, The College Board Review, No 113 (Fall 1979), 24-30.

Findley, Warren G., Carl C. Brigham Revisited. The College Board Review, No 119 (Spring 1981), 6-9.

Hecht, L. and Swineford, A., Item Analysis at the Educational Testing Service, Princeton, New Jersey: ETS, 1981.

Multiple Choice Questions: A Close Look. Princeton, N.J.: Educational Testing Service, 1973. (Publication #0077004 T40P5X. 252507)

TABLE 1

SUMMARY OF QUESTION SLOPE TYPES*
AND THEIR CHARACTERISTICS

SLOPE TYPE	MEAN % PASSING	% OF QUESTIONS	POINT BISERIAL COEFFICIENT RANGE	% CONCEPTUAL/ % FACTUAL
LINEAR	69.5%	19.0%	.1 TO .7	45/55
CEILING	89.7%	8.5%	-.4 TO .7	55/45
HIGH-END DISCRIMINATOR	62.3%	10.5%	.1 TO .9	38/62
FLAT LOW	28.5%	2.0%	-.6 TO .5	61/39

* ON ITEM/TEST REGRESSION CURVES.

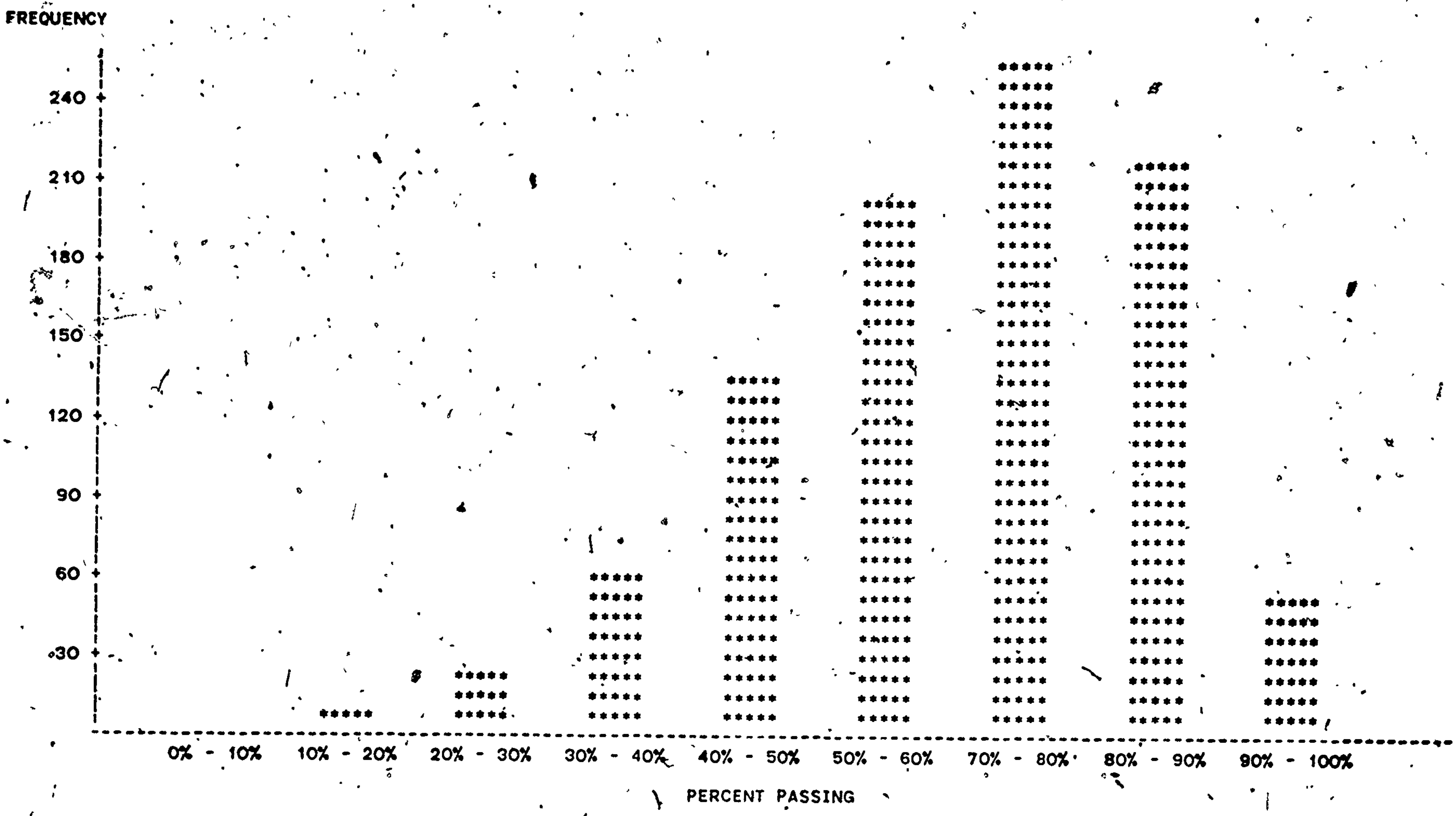

Figure 1. Frequency of test questions grouped by percent passing.

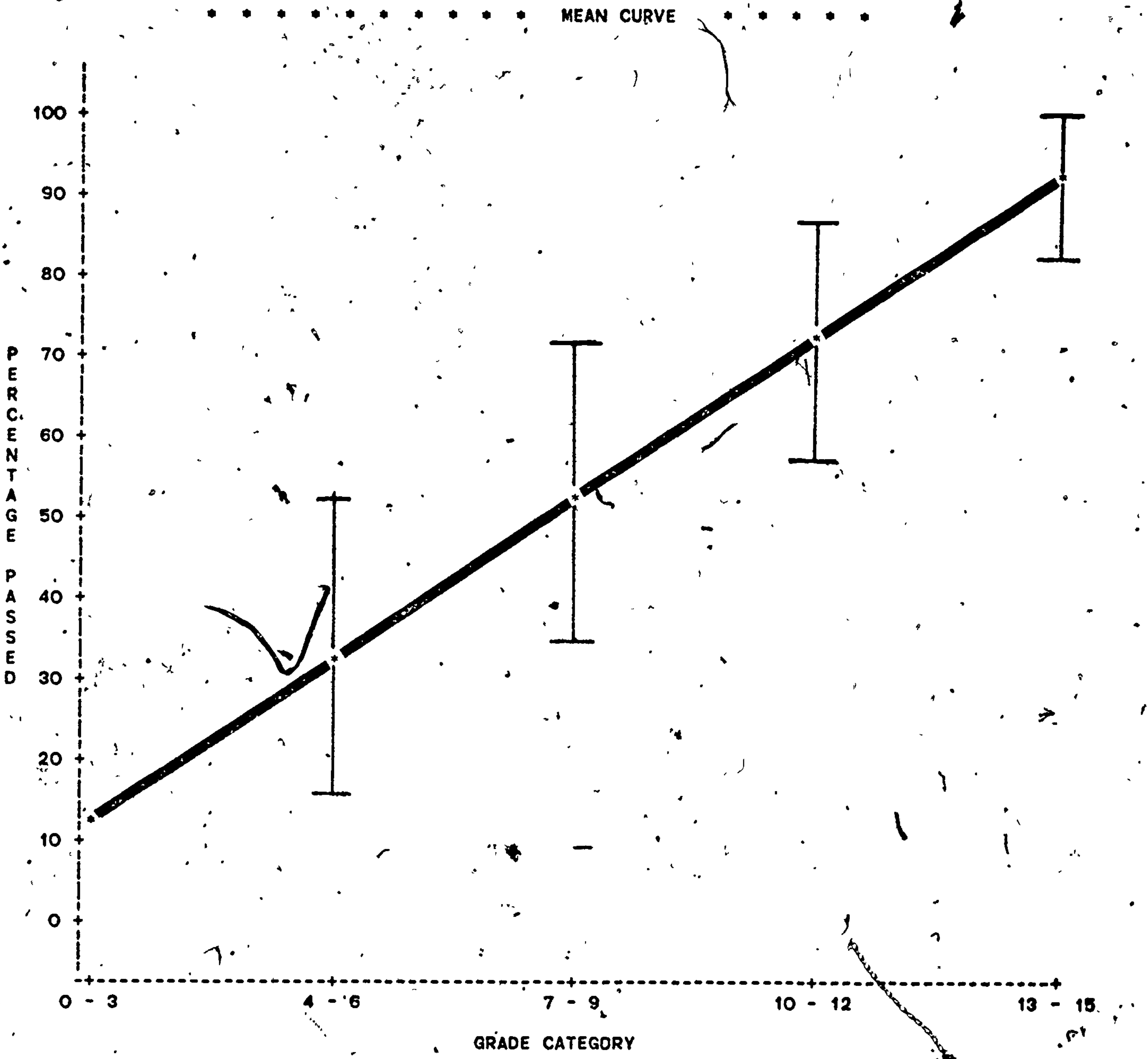

Figure 2. Item/test regression line for 945 test items in introductory Psychology.

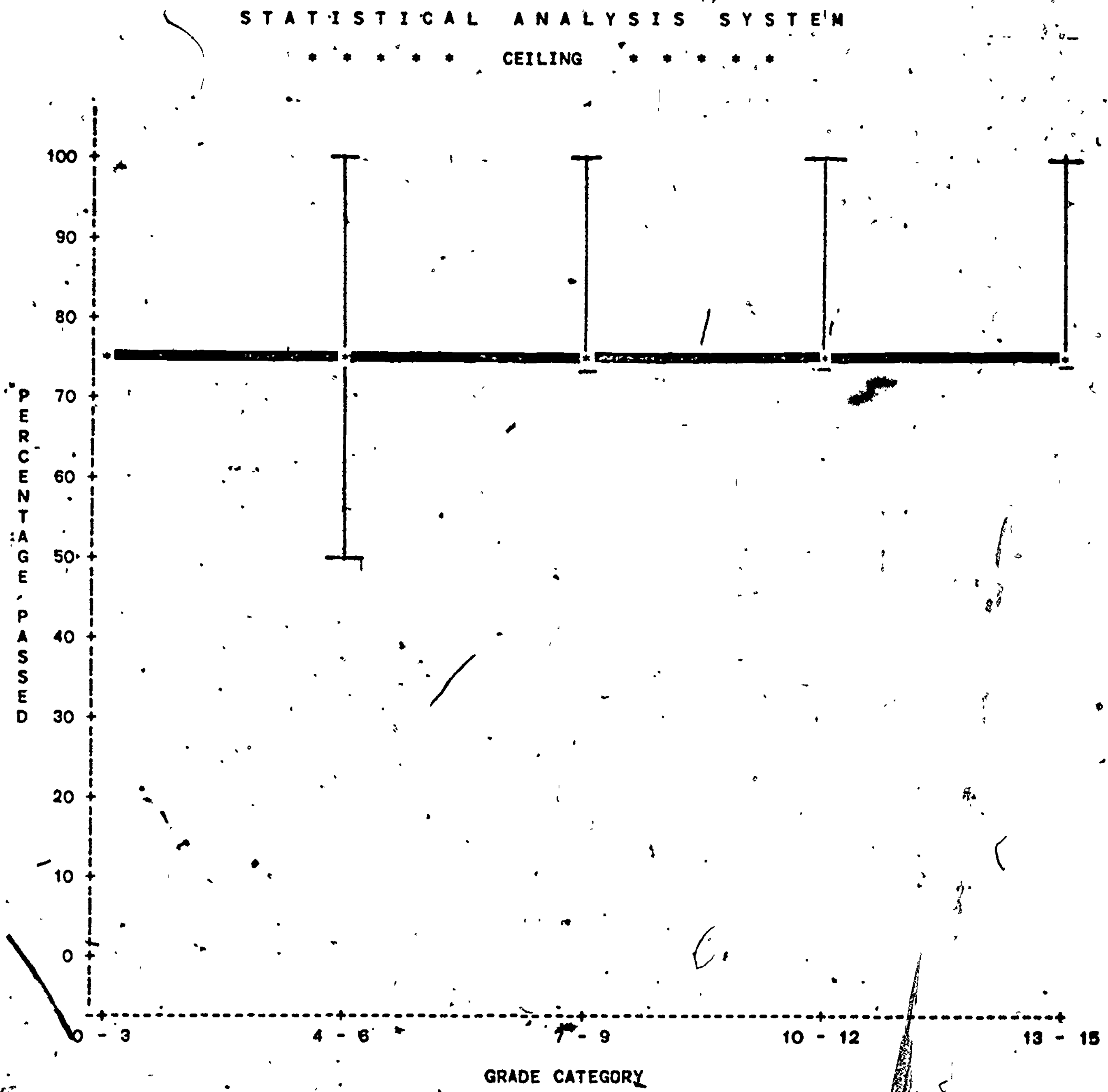

Figure 3. Sample item/test regression line for questions having a ceiling effect (flat-high slope).

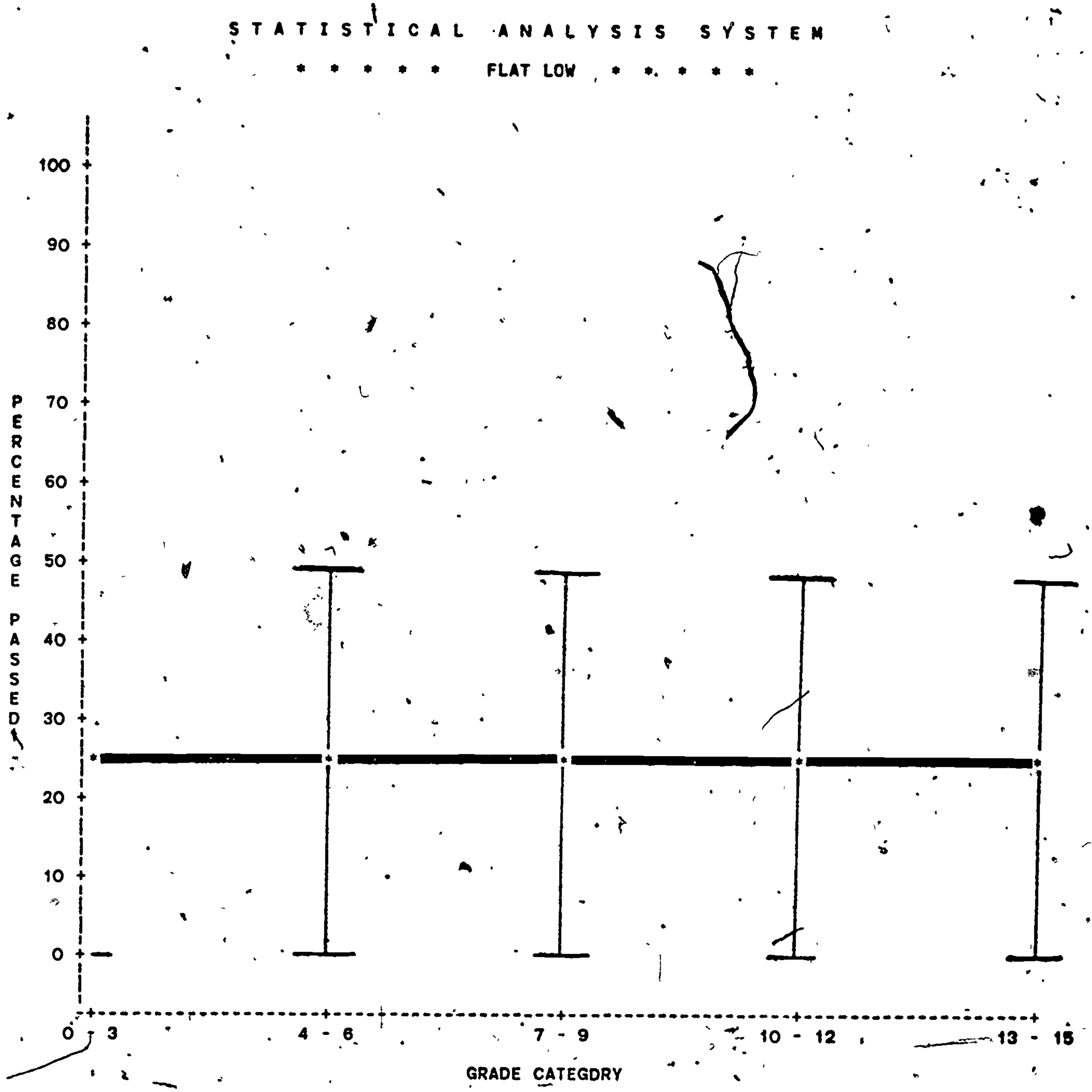

Figure 4. Sample item/test regression line for questions having flat-low slopes.

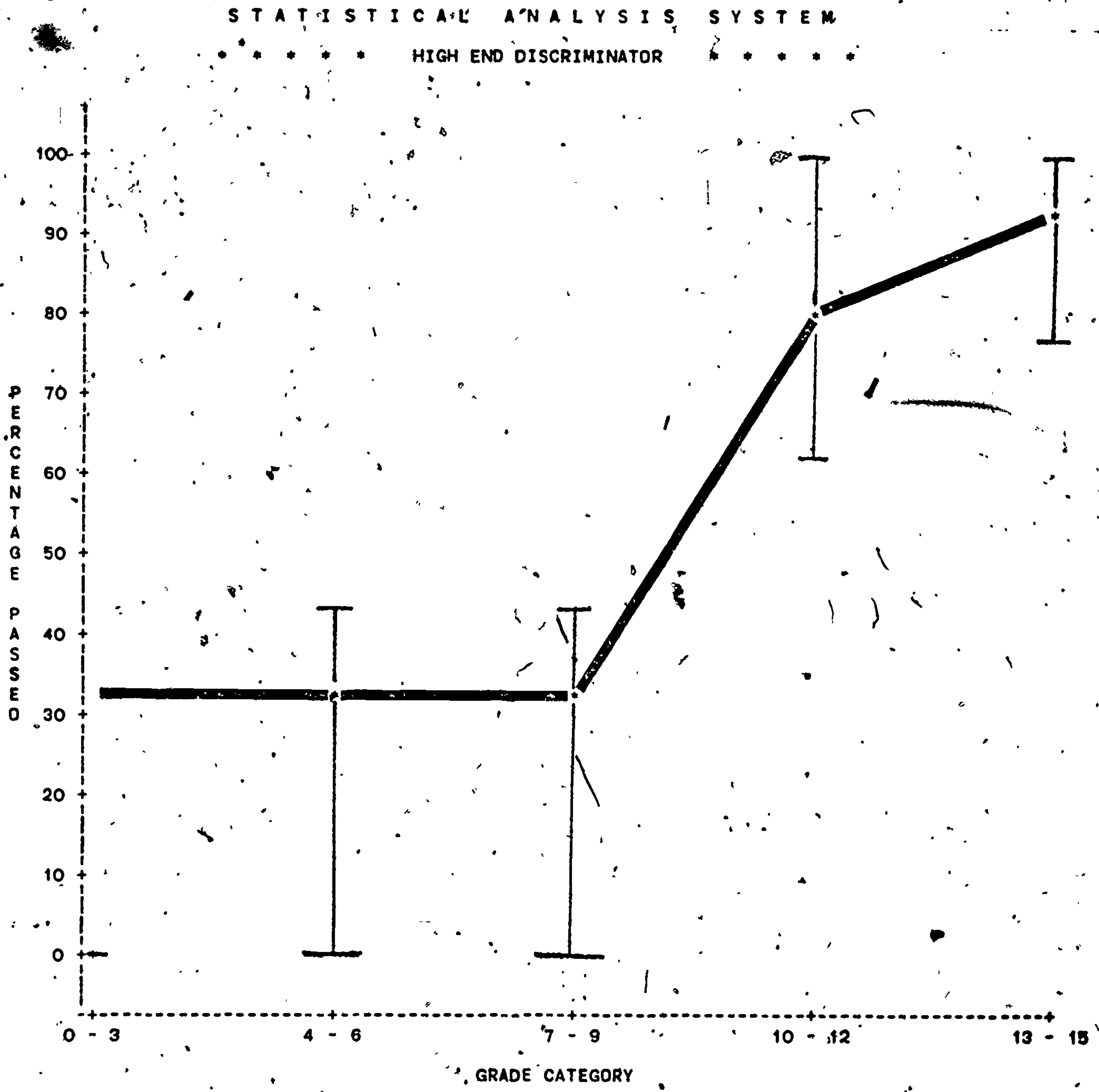

Figure 5. Sample item/test regression line for questions categorized as "high-end-discriminators.

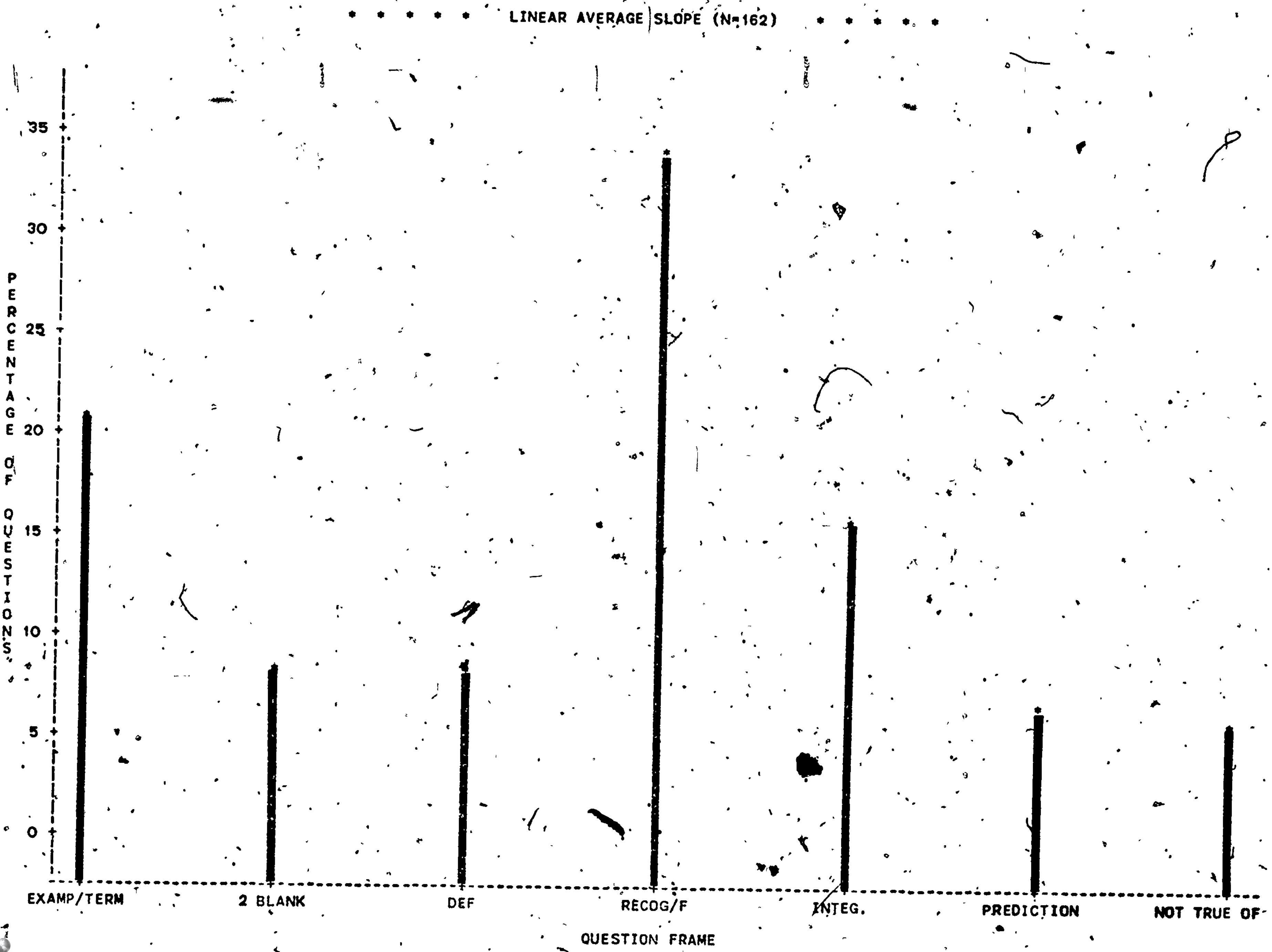

Figure 6. Distribution of question frame categories found for test questions producing linear slopes on item/test regression lines.

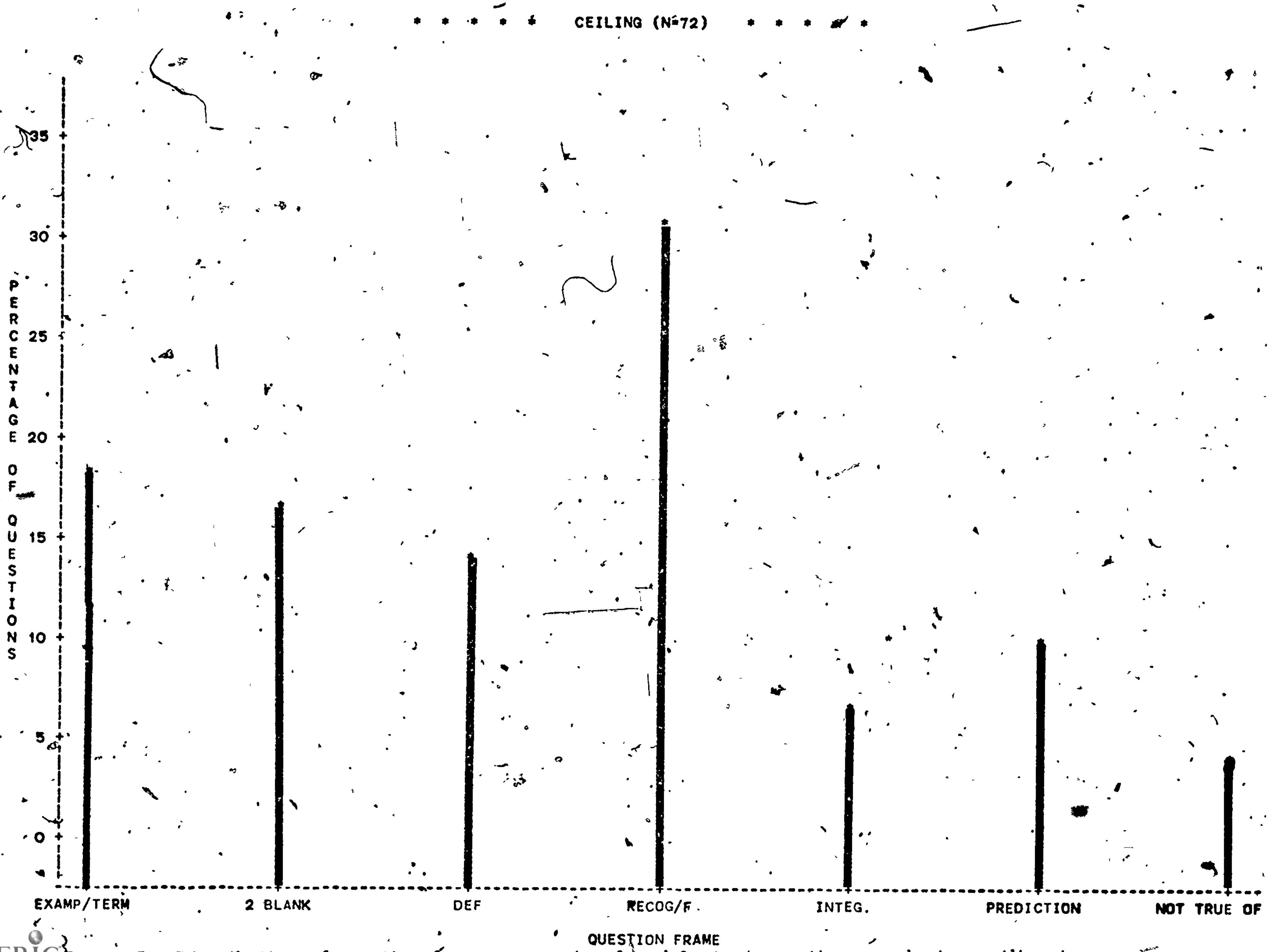

Figure 7. Distribution of question frame categories found for test questions producing ceiling-type slopes on item/test regression lines.

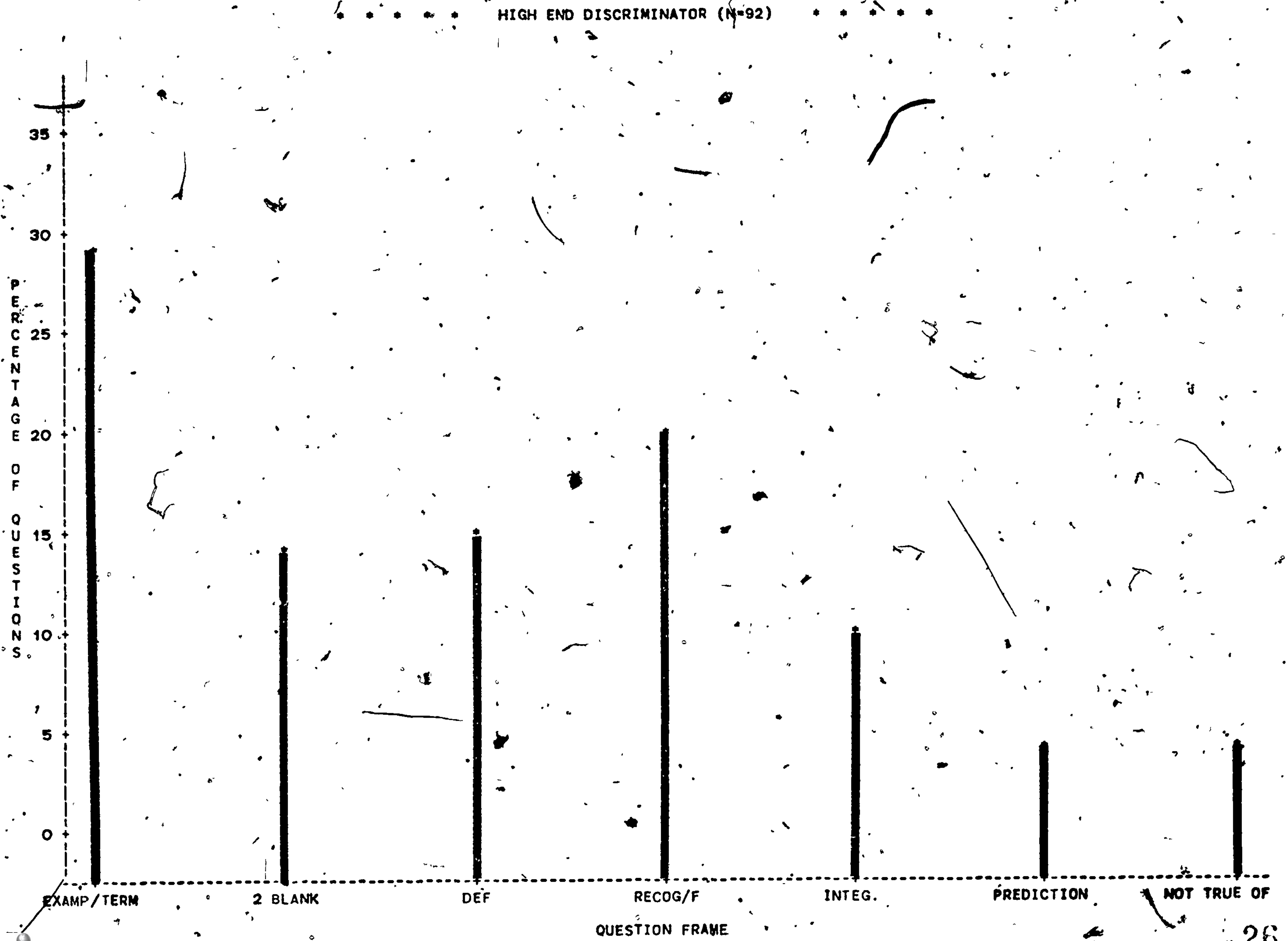

Figure 8. Distribution of question frame categories found for test questions producing "high-end-discriminator" type slopes on item/test regression lines.