DETERMINANTS OF ITEM DIFFICULTY:

A PRELIMINARY INVESTIGATION

Jason Millman

CSE Report No. 114
July 1978

Center for the Study of Evaluation
UCLA Graduate School of Education
Los Angeles, California 90024

The research reported herein was supported in whole or in part by the National Institute of Education, Department of Health, Education, and Welfare. However, the opinions expressed herein do not necessarily reflect the position or policy of the National Institute of Education, and no official endorsement by the National Institute of Education should be inferred.

Determinants of Item Difficulty:

A Preliminary Investigation

Jason Millman[1]
Cornell University

Test items, all referencing the same instructional objective, are not equally difficult. Some are hard with few students answering them correctly; others are easy with practically all students getting them right. Even seemingly subtle changes in the item content or how questions are asked (format changes) may result in substantial changes in the percent of students who can pass the item. What are the content and format characteristics that distinguish difficult from easy questions? This investigation attempts to identify some of the determinants of item difficulty within the context of a first course in educational statistics.

Significance of the Work

Item difficulty is only a secondary concern with norm-referenced tests. Although extremely hard or easy items are not desired on norm-referenced tests, the primary consideration is item discrimination. Discriminating items are needed to make comparative judgements. Criterion- or domain-referenced test interpretations, however, attempt to be absolute rather than comparative. Statements about the proportion of test items a student can answer correctly depend directly upon the difficulty of the items chosen to be on the test.

---

[1]This investigation was performed while the author was a Visiting Scholar at the Center for the Study of Evaluation. Special appreciation goes to W. Scott Outlaw who, with Center support, provided much of the computer programming work.

There is frequently a near infinite number of ways that mastery of an objective can be assessed. Ideally, to insure a proper inference about a student's ability with respect to the given skills or knowledges, the examiner would want to include items that represent the myriad assessment variations. Limited testing time permits the selection of only a few items. Knowledge about the determinants of item difficulty would aid test makers to select those content and format variations they should include in their sample of items to be reasonably sure that the skill or knowledge domain has been sampled broadly.

Knowledge about the determinants of item difficulty can also be of assistance to instructors and curriculum developers. Such information can be a guide in the choice of which discriminations and practice opportunities should be given most weight in the learning materials.

The Data Base

The author has developed a system for generating by computer multiple variations of each item that might appear on a test. Stored in the computer are item programs rather than specific items, as is done with most traditional computer-generated test construction systems. Each item program, written in a specially constructed language that facilitates the writing task, is capable of producing many versions of an item form. Shown in Figure 1 is an example of item form and two of the possible items that can be generated from it.[2]

---

[2]Further information about the system may be found in Jason Millman and W. Scott Outlaw, Testing by Computer. Association for Educational Data Systems Journal, 1978, 11, 57-72.

Item Form C6007

If {N} scores having a mean of {M} and a standard deviation of {S} were normally distributed, {how many / what percent} of the scores would be {less than / greater than} {X- / X+} ?

Notes:
N = RANDOM(10, 200, 10)
M = RANDOM(10, 100)
S = RANDOM(.2M, .3M)
X- = RANDOM(M-3S, M) X- ≠ M
X+ = RANDOM(M, M+3S) X+ ≠ M

Two possible test items

1. If 80 scores having a mean of 47 and a standard deviation of 10 were normally distributed, what percent of the scores would be less than 26?

2. If 140 scores having a mean of 75 and a standard deviation of 19 were normally distributed, how many of the scores would be greater than 71?

Figure 1. Example of an item form and some possible test items.

The brackets in the item form indicate places where substitutions can be made. As specified in the notes beneath the item form, various numbers could be substituted for N, M, S, X-, and X+. Further, some items (such as the second item shown in Figure 1, might ask "how many" and other items "what percent." Similarly, "less than" or "greater than" might be chosen. Finally, the score might be less than the mean (X-) or greater than the mean (X+).

The author wrote 133 item forms dealing with content taught in elementary statistics. Items generated from these forms appeared in a series of mastery tests administered as part of his course. Student answers were later typed by a secretary into the computer. Using a monitoring system created for the

purpose, it was possible to obtain item difficulty indices for each of the planned variations.

Figure 2 is a reproduction of the score tabulations for Item Form C6007, the form described in Figure 1. The trivial variations in the values of N, M, and S were not recorded in the computer printout. A score category of zero means no points deducted (answer was correct). Of the 23 times that "how many" appeared in the item, 12 times (or 52% of the time) the question was answered correctly. In contrast, the questions asking "what percent" were answered correctly 93% of the time. The versions "z neg" and "z pos" reference whether the "scores" referred to in the item were less than the mean (X-) or greater than it (X+).

Item Form C6007

--Version--		Score Category Frequencies									%
No.	Value	0	1	2	3	4	5	6	7	8	Total	Correct
1	how many	12	1	4	2	0	4	0	0	0	23	52
1	what percent	13	0	0	0	0	1	0	0	0	14	93
2	greater than	13	1	2	1	0	1	0	0	0	18	72
2	less than	12	0	2	1	0	4	0	0	0	19	63
3	z neg	14	0	1	2	0	1	0	0	0	18	78
3	z pos	11	1	3	0	0	4	0	0	0	19	58
Totals		75	3	12	6	0	15	0	0	0	111	68

Figure 2. Score tabulations for Item Form C6007.

Limitations of the Data

The fact that the entire data base consists of responses from one class of students seriously limits the generalizability of the results. The importance of a factor can obviously be influenced by the instruction to which the students were exposed. For example, the items illustrated in Figure 1 require the student to look up a percent in a table of areas under the normal curve. Had the author emphasized in his classroom instruction the additional step of converting these percents to frequencies (as required when the "how many" version appears), it is doubtful that the item difficulty of this version (52%) would have differed so much from 93%, the percent of students who answered the question correctly when the conversion to frequencies was not needed.

Nevertheless, although observed differences in item difficulties among the versions can be reduced by appropriate instruction, their existence indicates the importance that the dimension on which the versions differ be considered in the test construction item sampling plan.

Relative item difficulties are also influenced by the ability of the student group. Whether or not the test items required calculations with decimal numbers was, with one exception, essentially unrelated to how hard the item was. With a less able group than the University students involved in this study, use of decimal numbers could have made the problems significantly harder.

Also important is the degree to which variation along a dimension or factor is permitted. Versions that differ only slightly on a factor are less likely to yield meaningful differences in item difficulty than versions in which

extreme levels are permitted to appear. For example, size of data display was found not to be meaningfully related to item difficulty. However, the versions being compared were rather close. In one case, a 2X3 matrix of numbers was compared with a 3X4 matrix. Had the comparison been between 2X3 and 50X100 matrices, a different result might have been found.[3]

Two additional limitations of the data base deserve to be mentioned. The small number of students in the class (N=30) resulted in insufficient power to estimate accurately the various effects. This failing is perhaps not so serious for a study intended to suggest a procedure for inquiry into the determinants of item difficulty as it would be for an investigation whose goal is to produce more definitive information about such relationships.

Also hindering the discovery of the determinants of item difficulty is the fact that most items were answered correctly by most students. Content and format variations are most apt to make a difference when students have an intermediate amount of control over the subject matter and not when they have near mastery of it.

This study did take advantage of a natural situation in which students were being instructed and tested in an on-going, semester-long course. Although the variables manipulated during generation of the items were not of particular theoretical interest, the procedure does allow not only more interesting variables to be considered in future studies but also permits the

---

[3]The problem is that of interpreting "variance components" in fixed effects designs. For further discussion, see G. V Glass and R. Hakstian, Measures of association in comparative experiments. American Educational Research Journal, 1969, 6, 403-414.

investigation of a large number of variants of item form and content in a single research effort.

Analysis Procedures

Most factors were treated as dichotomies and corresponding 2x2 contingency tables were set up. Phi coefficients and the associated chi square statistic using a correction for continuity were computed. These data are illustrated in Table 1 for the how many/what percent factors. (See Figure 2 for the raw data.)

Table 1

Relationship between the How Many/What Percent Factor and Item Difficulty

	Factor: How Many	Factor: What percent	Total	
Right	12	13	25	ϕ = -.42
Wrong	11	1	12	χ^2 = 4.85
% correct	52%	93%	68%	p = <.03
				Estimated Difference (80%)=.34

In the example shown in Table 1, the correlation, ϕ, is -.42 indicating a moderately strong relationship between the factor and whether or not the item is answered correctly. This relationship is significant at the 3% level (2-tail).

In this example, the item was answered correctly 25 out of 37 or 68% of the time. The difference in item difficulty was .41 (that is .93 - .52 = .41). To facilitate comparisons among the factors studied, this difference in item difficulties was estimated for the situation in which the item is answered correctly 80% of the time (in contrast to 68% in this case) and in which the two versions are presented equally often. Thus, had the contingency table had 80% of the cases in the "Right" row, equal number of cases in each column, and $\phi = -.42$, then the difference in item difficulties would have been .34 (instead of .41). Unless the "p" value is low, such differences are subject to wide sampling fluctuations and should be so interpreted.

Frequently available are several contingency tables of data representing different items yet involving the same factor. With such replications, the several ϕ coefficients are averaged (using a Fisher Z transformation) and the difference in item difficulties assuming an 80% overall success level is estimated using this mean value.

The overall significance level is computed by summing the z-score equivalents of the χ^2 values and dividing by the square root of the number of data sets.[4] Since the same examinees frequently contribute more than one data value, none of the probability values shown in the tables to follow should be taken too seriously. Further, the χ^2 values are only approximations because sample sizes are frequently small, although all data sets that did not contain at least four observations in the "Wrong" row were eliminated.

[4]Richard B. Darlington, Radicals and squares: Statistical methods for the behavioral sciences. Ithaca, N. Y.: Logan Hill Press, 1975, p. 525.

Findings

Many of the results are presented in Tables 2 through 6. An attempt has been made to divide the item variations into homogeneous groupings.

The effect of changes in format on item difficulty is shown in Table 2. For the 23 true-false item forms considered in Table 2, there was but a minor tendency to answer "true" statements correctly more often than "false" statements. The mean correlation is .07, and the probability of such a value if the true relationship were zero is high (about .33). On an item in which 80% of the students answer correctly, this format variation would be expected to make a .05 difference in difficulty level.

None of the format variations resulted in particularly strong relationships with item difficulty. The possible exception is whether or not the right answer is included among the options of "none-of-the-above" questions. Sampling error is so large that an accurate estimate of its effect is not possible. Not shown in Table 2 is a significant ($p < .01$) tendency for students to identify statements that are sometimes true as "never" true rather than "always" true.

Linguistic variations are related to item difficulty in Table 3. None of the relationships is significant at a high level, although the direction of the relationships makes sense in several instances. Items containing the word "transform" are easier. Such a term was used in the textbook to signal problems of the type being tested for. Problems containing "X" and "Y", the symbols used in the corresponding formulas in the textbook, were slightly easier than problems not containing such symbols.

The word order, in all three cases, had a nonsignificant relation with item difficulty. Not identified in Table 3 are six item forms each of which

Table 2

Relationship between Item Format and Item Difficulty

Type of Question	Specific Description[a]	No. of ϕ's	Mean ϕ	Prob.	Difference[b]
True-False	True statements vs. False statements	23	.07	<.33	.05
Matching	Column 1 is ordered. Column 2 is: ordered vs. not ordered	2	.08	>.50	.06
None-of-the-above	Right answer is not among options vs. Right answer is given	1	.27	<.30	.22
Data Display	More numbers to read vs.Fewer numbers	4	.02	>.50	.02

[a]The easier version is presented first.

[b]Estimated difference in item difficulties between the easier and harder versions if 80% of the students on the average were to answer the questions correctly.

Table 3

Relationship between Language and Item Difficulty

Category	Specific Description[a]	ϕ	Prob.	Difference[b]
Synonymous Expressions	"convert" vs. "change"	.14	>.50	.12
	"transform" vs. "convert" or "change"	.21	<.39	.17
	X% are A vs. Probability of A = X	.08	>.50	.07
	Symbol (e.g., $_{.90}\chi^2_2$) vs. Verbal (eg. 90th percentile of a...)	.25	<.12	.20
Symbol Substitutions	"X" and "Y" vs. "U" and "V"	.15[c]	<.15	.12
	"c" or "k" vs. "a" and "b"	.06	>.50	.05
Word Order	S1 in a S2 vs. In a S2, S1	.11	>.50	.09
	Probability Y will be an X vs. Probability X will be a Y	.14	>.50	.11
	An individual..half of the time vs. Half of the time an individual..	.20	>.50	.16

[a]The easier version is presented first.

[b]Estimated difference in item difficulties between the easier and harder versions if 80% of the students on the average were to answer the questions correctly.

[c]Mean of four phi coefficients.

Table 4

Relationship between Variation in Calculation Requirements and Item Difficulty

Category	Specific Description[a]	No. of ϕ's	Mean ϕ	Prob.	Difference[b]
Negative numbers	Calculation involves all positive numbers vs. Some negative values	8	.08	<.23	.06
Decimal numbers	Calculations involving all inter numbers vs. Some decimal numbers	4	.09	>.50	.07
Arithmetical operations	Multiplication vs. Division	2	.01	>.50	.01
	Subtraction vs. Addition	2	.09	>.50	.07
Accuracy	Estimate vs. Calculate exactly	2	.23	≤.02	.18
Amount	More numbers in calculation vs. Fewer numbers	5	.01	≥.50	.01
Complexity	Fewer steps vs. More steps	6	.21	<.09	.17

[a]The easier version is presented first.

[b]Estimated difference in item difficulties between the easier and harder versions if 80% of the students on the average were to answer the questions correctly.

Table 5

Relationship between Variations in Questions about Areas Under Theoretical Distributions and Item Difficulty

Category	Specific Description[a]	No. of ϕ's	Mean ϕ	Prob.	Difference[b]
Area Under Normal Curve	Area to the right of a given z-score vs. Area to the left	2	.01	>.50	.01
	Area associated with a negative z-score vs. Area for a positive z	1	.21	<.35	.17
	Area < (or >) a single z-score vs. Area between two z's	1	.17	<.13	.13
	Area between two z-scores of different signs vs. Area when signs are the same	3	.17	<.15	.14
Significance Levels	1% level vs. 5% level	3	.05	>.50	.04
	99% confidence interval vs. 95% confidence interval	2	.17	<.46	.14
	1% or 5% level vs. 2% or 10% level	2	.00	>.50	.00
	2% level vs. 10% level	2	.28	<.14	.23

[a]The easier version is presented first.

[b]Estimated difference in item difficulties between the easier and harder versions if 80% of the students on the average were to answer the questions correctly.

Table 6

Relationship between Variations in Miscellaneous Questions and Item Difficulty

Category	Specific Description[a]	No. of ϕ's	Mean ϕ	Prob.	Difference[b]
Summation Operator	Simple sum vs. Sums of squares, products, etc.	3	.14	<.13	.11
	Sum of squares vs. Sum of products	2	.21	<.36	.17
	Sum of squares vs. Square of sums	2	.17	<.21	.14
Distributions	Skewed vs. Not skewed	3	.02	>.50	.02
	Shape of distribution identified vs. Shape not given	4	.39	<.01	.31
Transformations	+ or - constant to each observation vs. X or ÷ each observation by a constant	2	.39	<.29	.31
	Transformations of the form cX+k vs. The form (X+k)xc	2	.04	>.50	.03
	Single observation is changed vs. New observation is added	2	.43	<.01	.35

[a]The easier version is presented first.

[b]Estimated difference in item difficulties between the easier and harder versions if 80% of the students on the average were to answer the questions correctly.

contains two variations in the order of the required statistical operation. Examples include performing an r to Z transformation vs. a Z to r transformation and given a property to name the measurement scale vs. given a measurement scale to identify the corresponding property. Again, the relationship to item difficulty was negligible, with a mean probability value for the six item forms of close to .50.

Many of the statistics questions involved calculations. Summarized in Table 4 is the effect of variations in the nature of the calculations on item difficulty.

The type and number of data values had only a small relationship to item difficulty. An exception, not shown in Table 4, is a specific item form which employed percent values less than one (e.g., .0072%). Items containing such small percents were missed a relatively greater percent of the time (ϕ=.47, $p < .07$).

Although the sheer number of calculations did not affect item difficulty, the number of different steps in the calculation was associated with how often the item was missed. Further, students found it easier to estimate a statistic than to compute its value without error. (Next year when calculators will be permitted, the difference between the estimation and calculation versions may not be so large.)

Additional variations in item content and their correlations with item difficulty are shown in Tables 5 and 6. Two significant ($p<.01$) content variations are whether or not the shape of a distribution is specifically stated in a question and whether the effect on distribution statistics of <u>changing</u> or <u>adding</u> a new observation is asked for. In the former example, students would

tend to assume a distribution was normal (or symmetric) when asked questions about the characteristics of such a distribution or statistics computed therefrom.

It is not surprising that changes in item content (rather than in item format or in linguistic structure) are associated with the large correlations. Unfortunately, these variations are frequently specific to the item and do not lend themselves to general formulation.

Perhaps a general rule is that questions that begin to require different knowledges are most apt to have different difficulty levels. For example, one item form identifies the probability of X, the probability of Y, and the probability of both X and Y. The question asking for the probability of X or Y is much easier than that asking for the conditional probability of X given X or of Y given X (ϕ=.48, $p<.01$). The implication of this rule for the educator or test constructor is to underscore the importance of identifying the range of knowledges to be taught or tested.

In spite of the limitations of the data mentioned above, this preliminary investigation does add additional support to the view that item difficulties can be made to fluctuate by changes in how the questions are asked. Further study, more systematically designed and employing variables of more theoretical interest, is clearly needed.