DOCUMENT RESUME

ED 254 537 TM 850 046

AUTHOR Hutchinson, T. P.
TITLE Nonsense Items in Multiple Choice Tests.
PUB DATE Dec 84
NOTE 14p.; Paper presented at the Annual Meeting of the British Psychological Society (London, England, December 1984).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Foreign Countries; *Guessing (Tests); *Mathematical Models; *Multiple Choice Tests; Psychometrics; *Scoring Formulas; Test Construction; *Test Items; Test Theory
IDENTIFIERS England; *Nonsense Test Items; Partial Knowledge (Tests); Signal Detection Theory

ABSTRACT

One means of learning about the processes operating in a multiple choice test is to include some test items, called nonsense items, which have no correct answer. This paper compares two versions of a mathematical model of test performance to interpret test data that includes both genuine and nonsense items. One formula is based on the usual assumption that knowledge is all or none. The alternative formula incorporates the notion of partial knowledge adapted from signal detection theory. Results of a chemistry test taken by 407 subjects with four nonsense and 20 genuine items are used to compare the two formulas. A moderate correlation between the predictions and the findings indicate both model variations have some success, but the partial knowledge formula is the more accurate. (Author/BS)

Presented at the London Conference of the British Psychological Society,

December 1984

Nonsense Items in Multiple-Choice Tests

T P Hutchinson

Department of Statistics & Operational Research
Coventry (Lanchester) Polytechnic
Priory Street
Coventry CV1 5FB
England

Abstract

One means of learning about the processes operating in multiple-choice tests is to include among the test items some which have no correct answer among the alternatives available. If one makes the usual assumption that knowledge is all-or-none, a formula for predicting what proportion of these nonsense items are responded to may be derived. A test of chemistry taken by 407 subjects is used to compare this with an alternative formula that incorporates the notion of partial knowledge. This latter is found to be the more accurate.

1. Introduction

"The assessment of partial knowledge and the control of guessing behaviour have been two goals of measurement specialists since the introduction of the objective-test format. .... Related to the problem of tendencies to omit or guess at items and the problem of chance success under answer-every-item instructions is the evaluation of a correct or incorrect response to an objective item. The extent of a respondent's knowledge concerning item i cannot be accurately assessed using conventional scoring procedures; an incorrect response is not sufficient evidence to conclude nothing is known concerning the item. A correct response is insufficient evidence for the converse." (Hritz and Jacobs, 1970.)

One means sometimes used to gain more information about the processes operating in tests of ability (or aptitude, intelligence, knowledge, achievement, etc) is to include among the test items some which have no correct answer among the alternatives available. (But I should add that a subject's reaction to such items, i.e. whether he or she attempts them, probably provides little or no evidence about his or her ability, though it may do about aspects of personality. Including nonsense items has generally been done for research reasons, nct educational ones, though Granich (1931) suggested that announcing their inclusion was a means of reducing guessing.) In principle, we may note a distinction between nonsense items for which the question is meaningful but all alternatives listed are wrong, and those for which the question is meaningless and a correct answer does not exist; but reasons for choosing one or other type of nonsense item have usually not been given.

Inclusion of nonsense items dates back more than 50 years (English, 1928; Thelin and Scott, 1928; Brinkmeier and Keys, 1930; Granich, 1931). Recent uses include the studies of Bauer (1971), Bliss (1980), Crocker and Benson (1976), Cross and Frary (1977), Hritz and Jacobs (1970), Jacobs (1975), Lo and Slakter (1973), Miles (1973), Slakter (1967, 1969), Slakter et al (1970, 1971, 1975), Waller (1974), and Wu and Slakter (1978).

In order to interpret test data incorporating nonsense items, we need a mathematical model of performance on tests that will apply to genuine items, to nonsense items, and preferably to other variant forms of test also. In the absence of such a model, we would be in danger of proceeding as some of the references cited above did: correlating the proportion of nonsense items responded to with such quantities obtained from responses to genuine items as the proportion responded to, the proportion answered wrongly, the proportion answered correctly, and the increase in score when a forced response is obtained on items originally left unanswered. These quantities do not have a sound basis - a high proportion of wrong answers, for example, could be due to a high tendency to guess, but it could also be due to low ability. A general framework for integrating nonsense items with genuine ones will now be described. The predictions of two specific versions of it will then be compared with data. (The results have previously been reported in Frary and Hutchinson, 1982. A fuller description of the theoretical background and more discussion of, and references to, earlier work are included in the present paper.)

2. A description of partial knowledge analogous to the signal detection model of perception.

2.1 Preliminary. Dismissal of models that postulate specific mechanisms

Descriptions of subjects' reactions to some types of item may possess a degree of mechanistic realism. For instance, a subject may know something about a particular alternative answer that eliminates it from consideration. (Asked to indicate whether Paris or Rome is the capital of France, the correct answer is given if the subject knows that Rome is the capital of Italy.) As a second example, the product of 2½ and 3¼ may be known to lie between 7½ and 10, thus eliminating alternatives such as 5¼ and 11½, without the full details of multiplying fractions being known. I do not quite rule out the feasibility of tailoring theories of partial knowledge to fit specific types of item. But since most tests contain items of many types, and since what is usually wanted is a single score representing some form of general ability, I think a theory of general applicability is to be preferred, even if by its abstractness it loses mechanistic realism.

2.2 Distributions of mismatch

We shall adapt our ideas from signal detection theory (Green and Swets, 1966). To explain errors when a subject is attempting to detect a faint stimulus, this supposes the subject responds according to whether the level of some internal sensation exceeds or falls below a threshold level; and that the sensation is variable (i.e. has some statistical distribution), both when the stimulus is presented and when it is not, the average levels being different in the two conditions. Similarly, we shall suppose that each alternative in each item generates within the subject a certain feeling of inappropriateness to the question posed. This feeling tends to be stronger for the incorrect alternatives than for the correct one, though there is appreciable random variation. The subject normally chooses the alternative that generated the lowest mismatch. But if all exceeded some threshold level, then the subject is unwilling to answer. (This threshold is naturally affected by the instructions given concerning guessing.)

2.3 Mathematical expression

Notation:

N = number of alternatives in each item,

c = proportion of items answered correctly,

w = proportion of items answered wrongly,

u = proportion of items not answered,

X represents the inappropriateness of an alternative. The greater the difference between its average levels for correct and for incorrect alternatives, the easier is the item (or the cleverer is the subject). Denote the distributions of X under the two conditions by F and G:
Probability of X exceeding the value x for correct alternatives = $F(x)$,
Probability of X exceeding x for incorrect alternatives = $G(x)$,
$F(x)$ being less than $G(x)$.

T represents the response threshold, such that if the inappropriateness level exceeds this for all the alternative answers to an item, no choice is made.

We can now write down equations for u, c, and w in terms of F and G:

$$u = F(T)\ [G(T)]^{N-1}$$

$$c = \int_{-\infty}^{T} \frac{-dF(x)}{dx} [G(x)]^{N-1}\ dx$$

$$w = 1 - u - c$$

What the second of these equations is saying is that the probability of the inappropriateness generated by the correct alternative taking a value x is the probability density corresponding to $F(x)$, $\frac{-dF(x)}{dx}$; that the probability of all the N-1 incorrect alternatives having higher levels of inappropriateness is $[G(x)]^{N-1}$; that the probability of both these things being true is the product of the probabilities; and, finally, we need to consider all possible values of x less than T, so we sum with T being the limit of integration.

For a given item, ability is measured by how different F and G are. So choose them so as to jointly contain a single parameter characterising ability, λ, and obtain λ in terms of c and w by eliminating T from the above equations. Some examples will show how this is done. (Further implications of this model of performance are described in Hutchinson, 1982.)

2.4 Specific examples

Some choices of F and G give rise to a simple expression for the ability parameter λ.

Firstly, for $0 < x < 1$ and $\alpha \geq 1$, let

$$\left.\begin{aligned} F(x) &= 1 - x \\ G(x) &= \begin{cases} \lambda(1-x)^{1/\alpha} & (1-\lambda^{-\alpha}<x<1), \\ 1 & (0<x<1-\lambda^{-\alpha}). \end{cases} \end{aligned}\right\} \quad (1)$$

In this case, $1-\lambda^{-\alpha} = c-\alpha w/(N-1)$, so that, since the left-hand side of the equation is an increasing function of λ, we have derived the general linear correction for guessing, each correct answer receiving 1 mark and each wrong answer receiving $-\alpha/(N-1)$ marks. The conventional formula is obtained by setting $\alpha = 1$.

Secondly, suppose that for $x > 0$,

$$\left.\begin{aligned} F(x) &= \exp(-x) \\ G(x) &= \exp(-x/\lambda) \end{aligned}\right\} \quad (2)$$

Then $\lambda/(N-1) = c/w$. If $F(x) = 1 - x$ and $G(x) = (1 - x)^{1/\lambda}$, the same equation is obtained, illustrating that a particular formula for λ does not imply a unique pair of functions F and G.

2.5 Implications for nonsense items

The most obvious assumption to make in the light of the above theory is that the probability distribution of mismatch for the alternatives given for the nonsense items is the same as that for the incorrect alternatives in the genuine items. Then the probability of leaving this nonsense item answered is $[G(T)]^N$, in which case the probability of giving an answer (a) is $a = 1 - [G(T)]^N$.

If equations (1) hold, then in the special case $\alpha = 1$

$$a = \frac{Nw}{(N-1)u + Nw} \quad (3)$$

(we have assumed that all items are sufficiently difficult that all subjects have a non-zero probability of giving a wrong answer), or, more generally

$$a = 1 - \left(\frac{(N-1)u}{(N-1)u + (\alpha+N-1)w}\right)^{N/(\alpha+N-1)} \quad (4)$$

In the case $\alpha = 1$, equations (1) are equivalent to the conventional model that each subject knows the answer with probability p (a measure of ability) and decides to guess with probability g if he or she does not, in which case the probability of being correct is $1/N$. Then for nonsense items, the probability of response will presumably be g. It turns out that g is given by (3) (Ziller, 1957).

If equations (2) hold,

$$a = 1 - u^{Nw/[(N-1)(1-u)]} \tag{5}$$

3. Comparison of theories with data

3.1 The dataset

Cross and Frary (1977) report the administration of a 4-alternative 20-item test of chemistry to 407 subjects. As well as the 20 genuine items, there were four nonsense items included. These were of the type in which the question was meaningless. The directions to the subjects were designed to encourage guessing but discourage wild guessing : "Your score will be the number of items you mark correctly minus a fraction of the number you mark incorrectly. You should answer questions even when you are not sure your answers are correct. This is especially true if you can eliminate one or more choices as incorrect or have a hunch or feeling about which choice is correct. However, it is better to omit an item than to guess wildly among all of the choices given."

3.2 Method for making comparison

Because each individual subject was exposed to only four nonsense items, the following procedure was adopted.

The subjects were grouped into ranges according to their value of expression (3). Then the mean proportion of nonsense items which were answered by the subjects in each group was found for comparison. Finally, the process was repeated with subjects being grouped according to their value of expression (5), rather than expression (3).

3.3 Results

The results are shown in Table 1. It can be seen (i) that both variants of this theory have some success, in that there is a moderate correlation between the predictions and the findings, (ii) that both tend to overestimate, and (iii) that formula (5) appears to be slightly better than formula (3).

Another way of presenting the results is to calculate, on a subject by subject basis, the correlation between their actual proportion of nonsense items answered (which could take only the values 0, ¼, ½, ¾, 1, since there were only four such items) and the predicted proportions. This was found to be .46 in the case of expression (3) and .52 in the case of expression (5). Trying different values of α in (4), a maximum correlation of .49 was obtainable for $\alpha = 8$. (It may be objected that the theories say that a should take particular values for given u and w, not merely that a should be correlated with this value. But the possibility that the distractors

Table 1.

Comparison of two theories with data: Probability of response to nonsense items.

(a) Theory (1) with $\alpha = 1$			(b) Theory (2)		
Proportion of responses predicted by (3)	Number of subjects	Actual mean response probability	Proportion of responses predicted by (5)	Number of subjects	Actual mean response probability
.90 - 1.00	203	.84	.90 - 1.00	173	.88
.80 - .90	69	.67	.80 - .90	56	.68
.70 - .80	70	.56	.70 - .80	61	.66
.60 - .70	36	.49	.60 - .70	50	.61
.50 - .60	22	.48	.50 - .60	33	.50
.00 - .50	7	.29	.40 - .50	19	.42
			.00 - .40	15	.30
Mean = .86	407	Mean = .71	Mean = .80	407	Mean = .71

for the nonsense items may not be of the same attractiveness as the distractors for the genuine items suggests we should be interested in how high the correlations are, as well as in how low are the differences between predictions and empirical results.)

4. Discussion

Perhaps the largest body of work on this subject is by Slakter and colleagues. He uses the term "risk taking on objective examinations" (rtooe) to refer to the propensity to attempt nonsense items and to Ziller's index (equation (3)) for legitimate items. As far as I know, he has not considered any competitors to Ziller's index, such as the present equation (5). Slakter (1969) reports the administration to 636 subjects of 4 tests (language aptitude, mathematics aptitude, language achievement, mathematics achievement). These each included 10 nonsense questions embedded in 30 or 40 legitimate questions. Measures of rtooe were calculated from the nonsense items (proportion attempted) and from the legitimate items (by Ziller's method). Slakter found (i) these two measures positively correlated (average correlation, over 4 tests and 6 schools, was 0.74), (ii) rtooe appeared to be a general trait, in the sense that there was a positive correlation between different tests (correlation for nonsense items, averaged over 6 pairs of tests and 6 schools, was 0.74, and for legitimate items was 0.58). From this and other studies he concludes that rtooe is a feature of personality, and related to dominance-submission, maladjustment, vocational choice, curriculum choice, and perception of risk in military situations. Furthermore, examinees low in rtooe tend to be penalised on test score.

Certainly it is important to know whether the subject should be credited with any partial knowledge he or she has, and if we decide we do want to do this, how it should be done. For instance, the medical profession has, over the past twenty years, gone in for various forms of multiple choice questions in an enormous way. Numerous articles have appeared in Medical Education, Medical Teacher, and elsewhere arguing the pros and cons (e.g. Harden et al, 1976; Anderson, 1981). To an extent, the controversies are all about fine-tuning a system that by and large works pretty well. But concern for the few but not negligable number of examinees whose scores are substantially affected by test format, directions, etc. keeps interest hot.

So far as my own work is concerned, two ways of extending it are evident. Firstly, more datasets need to be examined to find whether they support equati[illegible] equation (5), or some other. Secondly, on the theoretical fro[illegible] choices of F and G should be compared with data: are there any others that lead to simple equations like (3) and (5)? If not, is it practicable to use some procedure like numerical integration? I would welcome news of any such studies.

References

Anderson, J. (1981). The MCQ controversy - A review. Medical Teacher, 3, 150-156.

Bauer, D.H. (1971). The effect of test instructions, text anxiety, defensiveness, and confidence in judgment on guessing behavior in multiple-choice test situations. Psychology in the Schools, 8, 208-215.

Bliss, L.B. (1980). A test of Lord's assumption regarding examinee guessing behavior on multiple-choice tests using elementary school students. Journal of Educational Measurement, 17, 147-153.

Brinkmeier, I. H. & Keys, N. (1930). Circumstantiality as a factor in guessing on true-false examinations. Journal of Educational Psychology, 21, 681-694.

Crocker, L. & Benson, J. (1976). Achievement, guessing, and risk-taking behavior under norm referenced and criterion referenced testing conditions. American Educational Research Journal, 13, 207-215.

Cross, L. H. & Frary, R.B. (1977). An empirical test of Lord's theoretical results regarding formula scoring of multiple-choice tests. Journal of Educational Measurement, 14, 313-321.

English, H.B. (1928). Bluffing in examinations. American Journal of Psychology, 40, 350.

Frary, R.B. & Hutchinson, T. P. (1982). Willingness to answer multiple-choice questions, as manifested both in genuine and in nonsense items. Educational and Psychological Measurement, 42, 815-821.

Granich, L. (1931). A technique for experimentation on guessing in objective tests. Journal of Educational Psychology, 22, 145-156.

Green, D. M. & Swets, J.A. (1966). Signal Detection Theory and Psychophysics. New York: Wiley.

Harden, R. McG., Brown, R.A., Biran, L.A., Dallas Ross, W.P., & Wakeford, R.E. (1976). Multiple choice questions: To guess or not to guess. Medical Education, 10, 27-32.

Hritz, R.J. & Jacobs, S.S. (1970). Risk taking and the assessment of partial knowledge. Proceedings of the 78th Annual Convention of the American Psychological Association, 171-172.

Hutchinson, T.P. (1982). Some theories of performance in multiple choice tests, and their implications for variants of the task. British Journal of Mathematical and Statistical Psychology, 35, 71-89.

Jacobs, S.S. (1975). Behavior on objective tests under theoretically adequate, inadequate and unspecified scoring rules. Journal of Educational Measurement, 12, 19-29.

Lo, M-Y. & Slakter, M.J. (1973). Risk taking and test-wiseness of Chinese students. Journal of Experimental Education, 42, 56-59.

Miles, J. (1973). Eliminating the guessing factor in the multiple choice test. Educational and Psychological Measurement, 33, 637-651.

Slakter, M.J. (1967). Risk taking on objective examinations. American Educational Research Journal, 4, 31-43.

Slakter, M.J. (1969). Generality of risk taking on objective examinations. Educational and Psychological Measurement, 29, 115-128.

Slakter, M.J., Crehan, K.D. & Koehler, R.A. (1975). Longitudinal studies of risk taking in objective examinations. Educational and Psychological Measurement, 35, 97-105.

Slakter, M.J., Koehler, R.A. & Hampton, S.H. (1970). Grade level, sex, and selected aspects of test-wiseness. Journal of Educational Measurement, 7, 119-122.

Slakter, M.J., Koehler, R.A., Hampton, S.H. & Grennell, R.L. (1971). Sex, grade level, and risk taking on objective examinations. Journal of Experimental Education, 39, 65-68.

Thelin, E. & Scott, P.C. (1928). An investigation of bluffing. American Journal of Psychology, 40, 613-619.

Waller, M.I. (1974). Estimating guessing tendency. Presented at the Annual Convention of the Psychometric Society and circulated as Research Bulletin 74-33 of the Educational Testing Service, Princeton,NJ.

Wu, T-H. & Slakter, M.J. (1978). Risk taking and test wiseness of Chinese students by grade level and residence area. Journal of Educational Research, 71, 167-170.

Ziller, R.C. (1957). A measure of the gambling response set in objective tests. Psychometrika, 22, 289-292.