דוח המרכז

252

The Point-Biserial as a Discrimination Index for Distractors in Multiple-Choice Items:

Deficiencies in Usage and an Alternative

Yigal Attali
Tamar Fraenkel

December 1998

מרכז ארצי לבחינות ולהערכה (ע"ר)
NATIONAL INSTITUTE FOR TESTING & EVALUATION
מיסודן של האוניברסיטאות בישראל
ת"ד 26015 ירושלים 91260, טל' 02-6759555

דוח מרכז 252
ISBN:965-502-027-4

The Point-Biserial as a Discrimination Index for Distractors in Multiple-Choice Items: Deficiencies in Usage and an Alternative

Yigal Attali and Tamar Fraenkel

December 1998

We thank Ruth Fortus, Yoav Cohen, and Naomi Gafni
for their time and helpful comments in reviewing earlier drafts of this report.

The Point-Biserial as a Discrimination Index for Distractors in Multiple-Choice Items: Deficiencies in Usage and an Alternative

Yigal Attali
Tamar Fraenkel
National Institute for Testing and Evaluation

We show that using the point-biserial as a discrimination index for distractors by differentiating between examinees who chose the distractor and examinees who did not choose the distractor is theoretically wrong and may lead to an incorrect rejection of items. We propose an alternative usage and present empirical evidence for its suitability.

Item discrimination refers to the degree to which an item differentiates correctly among examinees in the behavior that the test is designed to measure. When the test as a whole is to be evaluated by means of criterion-related validation, the items themselves may be evaluated and selected on the basis of their relationship to a criterion. In many achievement and aptitude tests, the criterion is the total test score. Since item responses are generally recorded as right or wrong, the measurement of item discrimination usually involves a dichotomous variable (performance on the item) and a continuous variable (performance on the criterion). Many different indexes of item discrimination have been developed and used, but, despite differences in procedures and assumptions, most of the indexes provide similar results (Oosterhof, 1976). In other words, although the numerical values of the indexes may differ, the items that are retained and those that are rejected on the basis of different discrimination indexes are largely the same.

The most frequently used classical index of item discrimination is some correlation measure between performance on the item and performance on the criterion. When performance on the item is dichotomous, the product-moment correlation is called the point-biserial correlation, *PB*. When the dichotomous variable creates a distinction between those examinees who answered correctly and those who did not answer correctly (PB_C), the formula is:

$$PB_C = \frac{M_C - M}{S}\sqrt{\frac{P_C}{1-P_C}}\ , \tag{1}$$

where M_C is the mean score on the criterion of the examinees who answered the item correctly; M and S are the mean and standard deviation on the criterion of all the examinees; and P_C is the proportion of the examinees who answered the item correctly.

When the items are multiple choice in format, it is important to study responses to distractors. Items may accidentally have more than one correct answer, or some distractors may perform inadequately, that is, the scores on the criterion of examinees who chose those distractors may be too high. Textbooks seldom give an in-depth treatment of distractor evaluation (Millman & Green, 1989), perhaps because not much is known about how to evaluate distractors (Haladyna, 1994).

A few alternative ways to analyze distractor performance can be found in the literature, for example, Wainer's (1989) trace lines, but for classical item analysis, the conventional method is to calculate *PB* for each distractor (PB_D). As for the correct answer, a dichotomous variable distinguishes between whether or not that distractor was selected. For a distractor, the formula which corresponds to formula (1) is:

$$PB_D = \frac{M_D - M}{S}\sqrt{\frac{P_D}{1-P_D}}\ , \tag{2}$$

where M_D is the mean score on the criterion of the examinees who selected the distractor, and P_D is the proportion of examinees selecting the distractor.

For most measurement purposes, an item which is deemed acceptable should have a positive PB_C, and each PB_D should be negative.

It is well known that correlation indexes are indexes of effect size. When the product-moment correlation is computed between a dichotomous variable and a continuous variable, namely, *PB*, it measures the effect size between the means of the two groups that are formed by the dichotomous variable. Another index of the effect size between the means of two groups is the *d* index of Cohen (1977), which is computed by dividing the difference between the means of the two groups by their joint standard deviation. *PB* and *d* are related by a straightforward formula:

$$PB = \frac{d}{\sqrt{d^2 + \frac{1}{P(1-P)}}}, \qquad (3)$$

where P is the proportion between one group and both groups.

It follows that the concept of item discrimination (measured by PB) is closely related to the concept of the effect size (measured by d) between the means of the two groups that are formed by the dichotomous variable used in calculating PB.

The main argument of this paper is that while PB_C is a valid and sound index for the assessment of the discrimination of the item as a whole, PB_D is wrong theoretically and biased statistically as an index for the assessment of the discrimination of the distractors. When PB_C is computed for the discrimination of the item as a whole, it forms a distinction between those examinees who answered correctly (i.e., chose the right answer) and those who did not answer correctly (i.e., chose one of the distractors). This is a meaningful distinction that is in accordance with the theoretical definition of item discrimination: It measures the degree to which the item differentiates among examinees in the behavior that the item is designed to measure - answering correctly or not. On the other hand, when PB_D is computed, a distinction is made between those examinees who chose the distractor (i.e., answered incorrectly), and those who did not choose the distractor (some of whom answered correctly and some of whom did not). This is not a meaningful distinction.

The distinction between examinees that is created by using PB_D as an index of discrimination implies the following: PB_D will receive negative values only when $M_D < M$ (see formula 2). However, there should be no a priori reason for a successful distractor to conform to this inequality. For example, if the item is a difficult one, it should make a distinction between examinees who have *high* criterion scores and examinees who have *medium* criterion scores. Thus, at least some of the distractors on this item should be chosen by examinees whose mean score is *average*, that is, around M. But such distractors would not, by definition, have a negative PB_D, even though M_D could be significantly lower than M_C - the mean score of the examinees who chose the correct answer.

A correct usage of PB for distractors is to change the two groups of examinees who are to be compared: One group should be those who chose the distractor, and the other group should be those

who chose the correct answer. The dichotomous variable that would be formed would make a meaningful distinction between those examinees who answered incorrectly *and* chose this specific distractor and between those examinees who chose the correct answer.

If we call this index PB_{DC} (for the comparison between Distractor and Correct choices), the formula for calculating it would be:

$$PB_{DC} = \frac{M_D - M_{DC}}{S_{DC}} \sqrt{\frac{P_D}{P_C}} \quad , \tag{4}$$

where M_{DC} is the mean score on the criterion of examinees who chose the distractor or the correct answer; S_{DC} is the standard deviation on the criterion of examinees who chose the distractor or the correct answer, and P_D and P_C are the proportion of examinees who selected the distractor and the correct answer, respectively.

What is the expected difference in the values of the two indexes, PB_D and PB_{DC} ? From their definitions it follows that PB_D will be too strict in its analysis of distractors and will tend to have higher (more positive) values than PB_{DC}. This is because PB_D is measuring the effect size of the difference between M_D and between the mean of the group of *all other examinees* (henceforth M_O), while PB_{DC} is measuring the effect size of the difference between M_D and between the mean of the group of *examinees who answered correctly* - M_C. M_O is almost always lower than M_C because it is composed of M_C and of the scores of other examinees who did not answer correctly. This means that the difference between M_D and M_O (used to calculate PB_D) will be smaller than the difference between M_D and M_C (used to calculate PB_{DC}). This could result in type II errors: Distractors whose M_D is significantly lower than M_C, but not significantly lower than M_O, will be incorrectly rejected.

In order to examine the magnitude of the difference in the values of the two indexes, PB_D and PB_{DC}, it is helpful to use the concept of threshold of acceptance and rejection of an item. As part of the process of selecting items, test developers have to set decision criteria for PB_C and PB_D. Items which do not meet these criteria do not enter the pool of operational items and may be subjected to some sort of revision. If, for example, the threshold for PB_D is set at -0.05, distractors with higher values will not be accepted. For distractors with a borderline PB_D value of -0.05, it is possible to

compute the corresponding value of PB_{DC}, if, in addition, the following characteristics of the item are known: The discrimination and difficulty of the item (PB_C and P_C), the proportion of examinees who chose the distractor (P_D), and the ratio of S_{DC} to S. For any given item characteristics and a borderline PB_D value for a distractor, the corresponding value of PB_{DC} will be the maximal value that will be accepted. Higher values of PB_{DC} will correspond to values of PB_D on the rejection side of the threshold. Table 1 presents several typical examples of item characteristics and the corresponding PB_{DC} for a distractor with a PB_D borderline value of -0.05. For simplification, the criterion scores are standard scores, and the only assumption in these calculations is that S and S_{DC} are equal.

TABLE 1

Maximal Values of PB_{DC} as a Function of Item and Distractor Characteristics

PB_C	P_C	PB_D	P_D	M_C	M_D	M_{DC}	PB_{DC}
0.3	0.7	-0.05	0.1	0.20	-0.15	0.15	-0.11
0.3	0.5	-0.05	0.2	0.30	-0.10	0.19	-0.18
0.3	0.3	-0.05	0.2	0.46	-0.10	0.23	-0.27
0.5	0.7	-0.05	0.1	0.33	-0.15	0.27	-0.16
0.5	0.5	-0.05	0.2	0.50	-0.10	0.33	-0.27
0.5	0.3	-0.05	0.2	0.76	-0.10	0.42	-0.42

Table 1 shows that item discrimination (PB_C) and item difficulty (P_C) are the two factors that influence the maximal value accepted for PB_{DC}: Greater discrimination and greater difficulty result in smaller values of PB_{DC} and in greater discrepancies between PB_D and PB_{DC}. When item difficulty is high (P_C equals 0.3, see third and sixth rows in Table 1), the maximal value of PB_{DC} is almost equal (in absolute value) to the discrimination value of the item itself! This is an absurd situation that may result in the rejection of many distractors even though the mean score of examinees who chose them is more than half a standard deviation lower than the mean score of examinees who chose the right answer. In other words, the power of PB_D to detect good distractors is very low for these items.

Another frequently used item discrimination index is the biserial (*BIS*), which is used when we assume that the latent variable underlying item performance is normally distributed. Is the discrepancy

between the two indexes, PB_D and PB_{DC}, different if their corresponding biserial indexes are used? Mathematically, the relationship between the biserial and the point-biserial is given by the formula:

$$BIS = \frac{\sqrt{P(1-P)}}{Y} PB \quad , \qquad (5)$$

where Y is the ordinate of the standard normal curve at the z-score associated with the P value for the item.

In most cases, applying formula 5 to PB_D and PB_{DC} will result in somewhat smaller discrepancies between their corresponding biserial indexes. For example, the discrepancies between the corresponding biserial of PB_D and PB_{DC} in Table 1 would be about 90% of their values for the point-biserial.

One difference between PB_D and PB_{DC} that must be taken into account is the fact that, compared to PB_D, PB_{DC} is always calculated on a smaller number of examinees: only those who chose the distractor or the correct answer. This results in a higher standard error for PB_{DC} relative to PB_D. If the number of examinees who answered the item is N, the standard error of PB_D is $1/\sqrt{N}$ and the standard error of PB_{DC} will be $1/\sqrt{(P_C + P_D)N}$. If the decision rule for distractors is intended to select one of the two hypotheses: H_0: $PB \geq 0$, or H_1: $PB < 0$ (although the meaning of the null value for PB_D and PB_{DC} is different because the two indexes measure different things), then the decision rule for PB_{DC} will have to be more extreme than the decision rule for PB_D, assuming that we want them to be technically (again, because they measure different things) equivalent. If DR is the decision rule for PB_D, then the decision rule for PB_{DC} will be $DR/\sqrt{P_C + P_D}$. The equivalence is in the sense that, under the assumption that PB_D or PB_{DC} are equal to null, the probability of having a sample PB_D or PB_{DC} that are lower than their respective decision rules, and thus committing a type I error, is equal.

Empirical Results

An empirical investigation was conducted to assess the effect of using PB_{DC} instead of PB_D in item evaluation. The investigation was conducted on items from the Psychometric Entrance Test (PET), a multiple-choice test which is constructed by the National Institute for Testing and Evaluation

(NITE). PET is used in the selection of students for admissions purposes by Israeli universities, and measures various scholastic abilities. We calculated PB_{DC} for 3620 items in the Verbal and Quantitative sections of PET that were piloted during 1997. Each pilot section is administered to approximately 300 examinees.

The purpose of the investigation was to assess the proportion of items that would be accepted or rejected by using PB_{DC} (with the adjusted decision rule discussed above) instead of PB_D, and the characteristics of these items in terms of difficulty and discrimination.

The items of interest were the items whose PB_C values were sufficiently high, but had been rejected because they had one or more distractors with an insufficiently low PB_D value. That is, the item discrimination was good, but the item had one or more distractors that did not meet NITE's standards ($PB_D \leq -0.05$). There were 523 such items, about 14% of the items studied. We assumed that if PB_{DC} were to affect decisions made about items, it would be with regard to these items, because the problematic distractors might, after all, have good discrimination values. Table 2 shows the number of items that were rejected because of PB_D values but had adequate PB_{DC} values, by levels of difficulty.

TABLE 2

Number of Items Rejected by Using PB_D, and Percentage of These Items Accepted by Using PB_{DC}, by Difficulty

Delta	Number of Items Rejected by Using PB_D	Percentage of These Items Accepted by Using PB_{DC}
<10	26	27%
10-11	37	30%
11-12	39	59%
12-13	49	71%
13-14	74	80%
14-15	83	89%
15-16	101	90%
16-17	59	93%
17<	55	100%
All	523	78%

Table 2 shows that almost 80% of the items rejected by using PB_D were accepted by using PB_{DC}. Moreover, the effect is stronger as the difficulty increases, with almost all of the very difficult items accepted. This effect of difficulty makes sense: Recall that the discrepancy between the indexes is greater as the difficulty of the item increases. The likelihood of finding distractors with mean score close to the overall item mean score is greater in difficult items.

The item discrimination of items that were rejected because of high PB_D values but had adequate PB_{DC} values is smaller than the item discrimination of all items that were accepted. The mean difference in the point-biserial index was 0.05, but in difficult items with a delta above 16 this difference was only 0.03. This result is also an expected one: The probability that an item will be rejected by using PB_D is greater as item discrimination decreases.

The final analysis conducted was the calculation of the distribution of the ratio of S_{DC} to S. For each of the items, S_{DC} was calculated for the candidate distractor for rejection, or in other words, for the distractor with the highest mean score of all the distractors. The median of the ratio of S_{DC} to S was 0.97. Only 15% of the items had a ratio greater than 1 (the minimum and maximum ratios were 0.85 and 1.18 respectively). Note that if the ratio is smaller than 1, the discrepancy between the two indexes, as measured in Table 1, will be even larger because the maximal values of PB_{DC} will be smaller.

Discussion

It is clear that the distractors in multiple-choice items are a very important part of the item, and can affect the pattern of responses to an item. It follows that, when the quality of a multiple-choice item is to be judged, it is important to analyze the performance of the item's distractors.

However, it is important to remember that the statistical analysis of distractors, for the purpose of accepting or rejecting the item, can be conducted only after it has been verified that the discrimination of the item as a whole has been satisfactory. A satisfactory discrimination index for the item means that the mean score of the examinees choosing the correct answer is significantly higher than the mean score of the examinees who chose one of the distractors. If this is found to be the case, then an analysis of each distractor can be conducted, whose purpose is similar to a post-hoc analysis:

Although the mean score of all the distractors is lower than the mean score of the correct answer, is this true of every distractor in itself? This defines a successful distractor as one whose mean score is significantly lower than the mean score of the correct answer.

An alternative definition of a successful distractor is that it should appeal to low-ability examinees and not to high-ability examinees. It should be remembered, however, that in the context of specific items, low and high abilities are relative to the overall difficulty of the item. In a difficult item, medium ability could be considered low relative to the ability of the examinees who answered the item correctly.

This paper showed that, when used conventionally, the point-biserial correlation as a discrimination index for distractors is not sensitive to the above considerations. It analyzes the difference between those examinees who chose the distractor and those who did not - some of whom answered correctly and some of whom chose other distractors. An alternative use of the point-biserial that does reflect the conceptual rational mentioned above was proposed. In this use, the difference between those examinees who chose the distractor and those who answered the item correctly is analyzed.

References

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences*. Florida: Academic Press.

Haladyna, T.M. (1994). *Developing and validating multiple-choice test items*. New Jersey: Lawrence Earlbaum.

Millman, J., & Green, J. (1989). The specification and development of tests of achievement and ability. In R.L. Linn (Ed.), *Educational measurement* (3rd Ed., pp. 335-366). New York: American Council on Education and Macmillan.

Oosterhof, A.C. (1976). Similarity of various item discrimination indices. *Journal of Educational Measurement, 13*, 145-150.

Wainer, H. (1989). The future of item analysis. *Journal of Educational Measurement, 26*, 191-208.