Shiken: JLT Testing & Evlution SIG Newsletter. 5 (3) October 2001 (pp. 13 - 17)

Statistics Corner: **Questions and answers about language testing statistics:**

Point-biserial correlation coefficients

James Dean Brown (University of Hawai'i at Manoa)

QUESTION: Recently on the email forum LTEST-L, there was a discussion about point-biserial correlation coefficients, and I was not familiar with this term. Could you explain what point-biserial correlation coefficients are and how they are important for language testers?

ANSWER: To adequately explain the point-biserial correlation coefficient, I will need to address four questions: (a) What is the point-biserial correlation coefficient? (b) How is the point-biserial correlation coefficient related to other correlation coefficients? (c) How is the point-biserial correlation coefficient calculated? And, (d) how is the point-biserial correlation coefficient used in language testing?

What Is the Point-Biserial Correlation Coefficient?

As I defined it in Brown (1988, p. 150), the point-biserial correlation coefficient (symbolized as r_{pbi}) is a statistic used to estimate the degree of relationship between a naturally occurring dichotomous nominal scale and an interval (or ratio) scale. For example, a researcher might want to investigate the degree of relationship between gender (that is, being male or female – a naturally occurring dichotomous nominal scale) and achievement in English as a second language as measured by scores on the end-of-the-year departmental examination (an interval scale).

Aside from the types of scales involved, the interpretation of the resulting coefficient is very similar to that for the more commonly reported Pearson product-moment correlation coefficient (sometimes referred to as Pearson *r*, or simply *r*). In brief like the Pearson *r*, the r_{pbi} can range from 0 to +1.00 if the two scales are related positively (that is, in the same direction) and from 0 to -1.00 if the two scales are related negatively (that is, in opposite directions). The higher the value of r_{pbi} (positive or negative), the stronger the relationship between the two variables. [For more detailed explanations of the interpretation and assumptions of Pearson r and r_{pbi}, see Brown, 1996, 1999.]

How Is the Point-Biserial Correlation Coefficient Related to Other Correlation Coefficients?

In distinguishing the point-biserial from other correlation coefficients, I must first point out that the point-biserial and biserial correlation coefficients are different. The biserial correlation coefficient (or *rbi*) is appropriate when you are interested in the degree of relationship between two interval (or ratio) scales but for some logical reason one of the two is more sensibly interpreted as an artificially created dichotomous nominal scale. For instance, you might be interested in determining the degree of relationship between passing or failing a first year university ESL course and language aptitude test scores. To do this, grades at the end of the course (A, B, C, D and F, often converted to a 4.00, 3.00, 2.00, 1.00, & 0.00 interval scale) might be artificially separated into a nominal scale made up of two groups: pass (A to D, or 1.00 to 4.00) and fail (F or 0.00). The degree of relationship between this new, artificially created dichotomy and the interval scores on the language aptitude test could then be determined by using the *rbi* coefficient. Thus the biserial correlation coefficient is appropriately applied when the nominal variable is artificially created (as in the pass-fail variable created from grade points), while the point-biserial correlation coefficient is appropriately applied when the nominal variable occurs naturally (as in the naturally occurring male-female gender distinction).

A variety of different correlation coefficients have been developed over the years for various combinations of scale types, as summarized in Table 1. The point-biserial is just one of these statistical tools (see the fifth row of correlation coefficients).

Table 1. ***Types of Correlation Coefficients***

Correlation Coefficient	Types of Scales
Pearson product-moment	Both scales interval (or ratio)
Spearman rank-order	Both scales ordinal
Phi	Both scales are naturally dichotomous (nominal)
Tetrachoric	Both scales are artificially dichotomous (nominal)
Point-biserial	One scale naturally dichotomous (nominal), one scale interval (or ratio)
Biserial	One scale artificially dichotomous (nominal), one scale interval (or ratio)
Gamma	One scale nominal, one scale ordinal

How Is the Point-Biserial Correlation Coefficient Calculated?

The data in Table 2 are set up with some obvious examples to illustrate the calculation of r_{pbi} between items on a test and total test scores. Notice that the items have been coded 1 for correct and 0 for incorrect (a natural dichotomy) and that the total scores in the last column are based on a total of 50 items (most of which are not shown).

Table 2. ***Example Student Data***

Table 2. *Example Student Data*

Student	Item 1	Item 2	Item 3	Item 4, 5, 6...	Total Score
Hachiko	1	0	1	. . .	50
Kazuko	1	0	1	. . .	45
Toshi	1	0	1	. . .	45
Yoshi	1	0	1	. . .	40
Tomoko	0	1	1	. . .	35
Yasuhiro	0	1	1	. . .	30
Yuichi	0	1	1	. . .	30
Masa	0	1	1	. . .	25
Mp	45	30	37.5	Total mean	37.50
Mq	30	45	.00	Standard Deviation	8.29
p	.50	.50	1.00		
q	.50	.50	.50	.00	
r_{pbi}	.91	-.91	.00		

To calculate the r_{pbi} for each item use the following formula:

$$r_{pbi} = \frac{M_p - M_q}{S_t}\sqrt{pq}$$

Where:

rpbi = point-biserial correlation coefficient

Mp = whole-test mean for students answering item correctly (i.e., those coded as 1s)

Mq = whole-test mean for students answering item incorrectly (i.e., those coded as 0s)

St = standard deviation for whole test

p = proportion of students answering correctly (i.e., those coded as 1s)

q = proportion of students answering incorrectly (i.e., those coded as 0s)

For example, let's apply the formula for r_{pbi} to the data for Item 1 in Table 2 (which we would expect to correlate highly with the total scores), where the whole-test mean for students answering correctly is 45; the whole-test mean for students answering incorrectly is 30; the standard deviation for the whole test is 8.29; the proportion of students answering correctly is .50; and the proportion answering incorrectly is .50.

$$r_{pbi} = \frac{M_p - M_q}{S_t}\sqrt{pq} = \frac{45-30}{8.29}\sqrt{(.50)(.50)} = \frac{15}{8.29}\sqrt{.2500} = 1.81(.50) = .91$$

Thus the correlation between item 1 and the total scores is a very high .91, and this item appears to be spreading the students out in very much the same way as the total scores are. In this sense, the point-biserial correlation coefficient indicates that item 1 discriminates well among the students in this group (at least in terms of the way the overall test discriminates).

As another example, let's apply the formula for r_{pbi} to the data for Item 2 in Table 2 (which we would expect to be highly but negatively correlated with the total scores), where the whole-test mean for students answering correctly is 30; the whole-test mean for students answering incorrectly is 45; the standard deviation for the whole test is still 8.29; the proportion of students answering correctly is still .50; and the proportion answering incorrectly is still .50.

$$r_{pbi} = \frac{M_p - M_q}{S_t}\sqrt{pq} = \frac{30-45}{8.29}\sqrt{(.50)(.50)} = \frac{-15}{8.29}\sqrt{.2500} = -1.81(.50) = -.91$$

Thus the correlation between item 2 and the total scores is a very high negative value of -.91, and this item appears to be spreading the students out opposite to the way the total scores are. In other words, the point-biserial correlation coefficient shows that item 2 discriminates in a very different way from the total scores at least for the students in this group.

As one last example, let's apply the formula for r_{pbi} to the data for Item 3 in Table 2 (which we would expect to have no correlation with the total scores), where the whole-test mean for students answering correctly is 37.5; the whole-test mean for students answering incorrectly is 0.00 because it is non-existent; the standard deviation for the whole test is still 8.29; the proportion of students answering correctly is 1.00; and the proportion answering incorrectly is .00.

$$r_{pbi} = \frac{M_p - M_q}{S_t}\sqrt{pq} = \frac{37.5 - .00}{8.29}\sqrt{(1.00)(.00)} = \frac{37.5}{8.29}\sqrt{.00} = 4.52(.00) = .00$$

Thus the correlation between item 3 and the total scores is zero, and this item does not appear to be spreading the students out in the same way as the total scores. In other words, item 3 is not discriminating at all among the students in this particular groupÑin this case because there is no variation in their answers.

How Is the Point-Biserial Correlation Coefficient Used in Language Testing?

As mentioned above, the point-biserial correlation coefficient can be used in any research where you are interested in understanding the degree of relationship between a naturally occurring nominal scale and an interval (or ratio) scale. For instance, I might be interested in the degree of relationship between being male or female and language aptitude as measured by scores on the Modern Language Aptitude Test (or MLAT; Carroll & Sapon, 1958). The point-biserial correlation coefficient could help you explore this or any other similar question. For examples of other uses for this statistic, see Guilford and Fruchter (1973).

However, language testers most commonly use r_{pbi} to calculate the item-total score correlation as another, more accurate, way of estimating item discrimination. The correlation coefficient being calculated here is between a naturally occurring dichotomous nominal scale (the correct or incorrect answer on each item usually coded as 1 or 0) with an interval scale test. Such item-total correlations are often used to estimate item discrimination. Consider the item analysis results shown in Table 3.

Item #	*IF*	r_{pbi}	* = p < .05
1	0.930	0.153	
2	0.656	0.295	*
3	0.882	0.122	
4	0.738	0.189	
5	0.455	0.310	*
6	0.838	0.394	*
7	0.684	0.469	*
8	0.552	0.231	
9	0.581	0.375	*
10	0.398	0.399	*
11	0.926	0.468	*
12	0.774	0.468	*
13	0.663	0.414	*
14	0.862	0.276	
15	0.624	0.205	

Table 3. ***Example Item Analysis*** *(for 32 students)*

The goal of the analysis shown in Table 3 is to estimate how difficult each item is (the *IF*, or item facility, shown in the second column) and how highly each item is correlated with the total scores (the r_{pbi} shown in the third column). The item facility, as estimated by the *IF*, ranges from 0.00 (everybody answered incorrectly) to 1.00 (everyone answered correctly) and shows how easy (or difficult) each item is. The r_{pbi} shows the degree to which each item is separating the better students on the whole test from the weaker students. Thus the higher the r_{pbi}, the better the item is discriminating. Notice in Table 3 that asterisks refer to the p > .05 at the bottom of the table and thereby indicate the items with point-biserial correlation coefficients that are significant at the .05 level (in other words, those items that have only a five percent chance of having occurred for chance reasons alone). [For more information on how to determine these *p* values for r_{pbi}, see Brown, 1996, p. 178; for more information on item analysis for norm-referenced testing purposes, see Brown, 1996 (pp. 64-74), or 2000a.]

Certainly, if you are interested in creating a shorter, more efficient, norm-referenced version of the test, you might be wise to select those items with the highest point-biserial correlation coefficients from among those that are significant (numbers 2, 5-7, & 9-13) for the new revised version of the test. At the same time, you should keep an eye on the item facility index shown in the first column so that you select a balance of items that average out to make a test that is neither to difficult nor too easy. This strategy is very similar to the way the discrimination index is used (for more on this statistic, see Brown, 1996, pp. 66-70). Such statistics can even be useful if what you need is a longer test: simply examine those items that appear to be discriminating well and write more items like them.

One important caveat: remember that item analysis statistics, like the r_{pbi}, are only tools that can help you in selecting the best items for a norm-referenced test, but they should never be used to replace the common sense notions involved in developing sound language test items. In other words, use these statistics to help you understand how students perform on your test items and then use that information to help you design a better test next time, while always keeping in mind your theoretical and practical reasons for writing the items you did and designing the test the way you did.

References

Brown, J. D. (1988). *Understanding research in second language learning: A teacher's guide to statistics and research design*. Cambridge: Cambridge University Press.

Brown, J. D. (1996). *Testing in language programs*. Upper Saddle River, NJ: Prentice Hall.

Brown, J. D. (trans. by M. Wada). (1999). *Gengo tesuto no kisochishiki*. [Basic knowledge of language testing]. Tokyo: Taishukan Shoten.

Brown, J. D. (2000a). Statistics Corner. Questions and answers about language testing statistics: How can we calculate item statistics for weighted items?. *Shiken: JALT Testing & Evaluation SIG Newsletter, 3* (2), 19-21. Also retrieved March 1, 2001 from the World Wide Web: http://jalt.org/test/bro_6.htm.

Brown, J. D. (2000b). Statistics Corner. Questions and answers about language testing statistics: What issues affect Likert-scale questionnaire formats?. *Shiken: JALT Testing & Evaluation SIG Newsletter, 4*(1), 18-21. Also retrieved March 1, 2001 from the World Wide Web: http://jalt.org/test/bro_7.htm.

Brown, J. D. (2001). *Using surveys in language programs*. Cambridge: Cambridge University Press.

Carroll, J. B., & Sapon, S. M. (1958). *Modern language aptitude test*. New York: The Psychological Corporation.

Guilford, J. P., & Fruchter, B. (1973). *Fundamental statistics in psychology and education*. (5th ed.). New York: McGraw-Hill.

HTML: http:// jlt.org/test/bro_12.htm / **PDF**: http:// jlt.org/test/PDF/Brown12.pdf