PRELIMINARY ITEM STATISTICS USING POINT-BISERIAL CORRELATION AND P-VALUES

BY
SEEMA VARMA, PH.D.

EDUCATIONAL DATA SYSTEMS, INC.
15850 CONCORD CIRCLE, SUITE A
MORGAN HILL CA 95037
WWW.EDDATA.COM

Overview

Educators in schools and districts routinely develop and administer classroom or benchmark tests throughout the academic year to get a quick overview of student mastery of relevant material. Generally these tests are used as diagnostic tools and are not used to measure growth over time or for official or statewide reporting purposes. For these reasons such tests are perceived to be low stakes and they undergo minimal or no statistical review for reliability and validity.

We argue that even though internally developed tests may appear to be less important due to their low-stakes nature, the fact that they are used for student diagnostics, as well as for classroom instruction, curriculum development, and to steer teacher professional development, makes them one of the most vital sources of information at the school and district levels. For these reasons, it is essential that the data these tests provide be reliable and meaningful.

We believe that with some pre-planning, schools and districts can significantly improve the quality of internally developed tests. One of the ways by which tests are checked for quality is analysis of each question or "item." The objective of this paper is to help educators conduct item analysis of internally developed assessments to obtain higher quality and more reliable test results.

What is Item Analysis?

Item analysis is a method of reviewing items on a test, both qualitatively and statistically, to ensure that they all meet minimum quality-control criteria. The difference between qualitative review and statistical analysis is that the former uses the expertise of content experts and test review boards to identify items that do not appear to meet minimum quality-control criteria. Such qualitative review is essential during item development when no data are available for quantitative analysis. A statistical analysis, such as item analysis, is conducted after items have been administered and real-world data are available for analysis. Statistical analysis also helps to identify items that may have slipped through item review boards; item defects can be hard to identify. The objective of qualitative and statistical review is the same – to identify problematic items on the test. Problematic items are also called "bad" or "misfitting" items. Items may be problematic due to one or more of the following reasons:

- Items may be poorly written causing students to be confused when responding to them.
- Graphs, pictures, diagrams or other information accompanying the items may not be clearly depicted or may actually be misleading.
- Items may not have a clear correct response, and a distractor could potentially qualify as the correct answer.
- Items may contain distractors that most students can see are obviously wrong, increasing the odds of student guessing the correct answer.
- Items may represent a different content area than that measured by the rest of the test (also known as multidimensionality).
- Bias for or against a gender, ethnic or other sub-group may be present in the item or the distractors.

Normally, output from a standard Rasch program, such as Winsteps, provides statistics on various item analysis indices. However, using such statistical programs and interpreting the output requires training. This document is intended to help educators in schools and district offices perform preliminary item analysis without extensive training in psychometrics. The item analyses we discuss here are *point-biserial correlations* and *p-values*. We will show how to compute and interpret these statistics using two different programs: Excel and SPSS.

Implications of Problematic Items in Terms of Test Scores

One may ask, why is it so important to review every item on a test? One may speculate that as long as the majority of the items on a test are good, there may not be much impact if a few items are problematic. However, based on statistical theory and previous experience, we know that the presence of even a few problematic items reduces overall test reliability and validity, sometimes markedly. All measurement tools, such as tests, surveys, and questionnaires, are assessed in terms of these two criteria, reliability and validity. We will briefly explain these two terms in the context of assessments.

Reliability tells us whether a test is likely to yield the same results if administered to the same group of test-takers multiple times. Another indication of reliability is that the test items should behave the same way with different populations of test-takers, by which is generally meant that the items should have approximately the same ranking when sorted by their p-values, which are indicators of "item difficulty." The items on a math test administered to fifth graders in Arizona should show similar p-value rankings when administered to fifth graders in California. The same test administered the next year to another cohort of fifth graders should again show similar rankings. Large fluctuations in the rank of item (one year it was the easiest item; the next year it was the hardest on the same test) would indicate a problem with the both the item and, by implication, the test itself.

Validity tells us whether the test is measuring what it purports to measure. For example, an algebra test is supposed to measure student competency in algebra. But if it includes word problems, such a test may be a challenge for students with poor English language skills. It would, in effect, be measuring not simply algebra skills but English language skills as well. Thus, the test would measure both algebra and English skills for one subgroup of students, algebra alone for the rest of the students for whom language is not an issue. This is different from the stated goal of being an *algebra* test for all students.

Both reliability and validity are important in any assessment. If a test is not reliable and valid, then the student scores obtained from that test are not reliable and valid and are therefore not indicative of the student's mastery of the material or his or her ability in that content. Non-reliable and non-valid test scores are simply meaningless numbers.

Thus, even when tests are administered at the school level (as opposed to a large-scale assessment), it is important to identify problematic items to ensure the test results are meaningful. While educators may reuse items over several years, it is advisable to remove bad items from the item pool. Besides lowering the reliability of the test, bad items also confuse students during the test-taking process. Students cannot be expected to be adept at identifying and rejecting items that appear incorrect and moving on to the next question. They spend time and energy responding to poorly written items.

Point-Biserial Correlation and P-Values

We now discuss two simple statistics used to determine whether a test item is likely to be valid and reliable: point-biserial correlations and p-values.

The *point-biserial correlation* is the correlation between the right/wrong scores that students receive on a given item and the total scores that the students receive when summing up their scores across the remaining items. It is a special type of correlation between a dichotomous variable (the multiple-choice item score which is right or wrong, 0 or 1) and a continuous variable (the total score on the test ranging from 0 to the maximum number of multiple-choice items on the test). As in all correlations, point-biserial values range from -1.0 to +1.0. A large positive point-biserial value indicates that students with high scores on the overall test are also getting the item right (which we would expect) and that students with low scores on the overall test are getting the item wrong (which we would also expect). A low point-biserial implies that students who get the item correct tend to do poorly on the overall test (which would indicate an anomaly) and that students who get the item wrong tend to do well on the test (also an anomaly).

The *p-value* of an item tells us the proportion of students that get the item correct. When multiplied by 100, the p-value converts to a percentage, which is the percentage of students that got the item correct. The p-value statistic ranges from 0 to 1.

Computation and Interpretation of Point-Biserial Correlation and P-Values

Illustrated below is a sample data matrix comprised of 10 items and 9 students (labeled Kid-A through Kid I). Items are represented in the matrix columns from left to right, and students are represented as rows. A value of "1" in the grid signifies that the student got the item correct; "0" indicates the student got it wrong. For instance Kid-A got all the items right except for Item 9; Kid-H got only Items 2 and 3 right and the rest of the items wrong.

Table 1: Sample Data Matrix

	Items									
Students	1	2	3	4	5	6	7	8	9	10
Kid-A	1	1	1	1	1	1	1	1	0	1
Kid-B	1	1	1	1	1	1	1	0	1	0
Kid-C	1	1	1	1	1	1	0	1	0	0
Kid-D	1	1	1	1	1	0	1	0	1	0
Kid-E	1	1	1	1	1	1	0	1	0	0
Kid-F	1	1	1	0	1	0	0	0	0	0
Kid-G	1	1	0	1	0	1	0	0	0	0
Kid-H	1	0	1	0	1	0	0	0	0	0
Kid-I	0	1	1	0	0	0	0	0	0	0

Computing Point-Biserial Correlations and P-Values in Excel

To compute point-biserials and p-values in Excel, replicate the sample data matrix, above, in an Excel worksheet. Now, referring to Table 2, the bold alphabetical letters on the top row (A, B, C, etc.) represent the column labels in Excel. Underneath them are the test item labels. Rows 1 through 10 label students Kid-A through Kid-I. The steps for computing the point-biserial correlation are as follows:

1. Compute the total student score (sum columns B through K for each row) as shown in Table 2, column L.

2. Compute the total score minus each item score, shown in Table 3, columns M through V, so that the total score minus the first item is in column M, the total score minus the second item is in column N, and so forth.

3. Compute the point-biserial correlation for each item using the "Correl" function. This computation results in the correlation of the item score and the total score minus that item score. For example, the Item 1 correlation is computed by correlating Columns B and M. To compute point-biserials, insert the Excel function =CORREL(array1, array2) into Row 12 for each Columns M through V, as shown in Table 4. The result is the point-biserial correlation for each item.

In Row 13, compute each item p-value by calculating the sum of the correct scores for each item, as shown in Table 2, Row 11, then divide this number by the total number of students who took that item (e.g., Item 1 has 8 correct answers and 9 students, or 8/9 = 0.89).

Table 2: Sample Student Data Matrix in Excel

	A	B	C	D	E	F	G	H	I	J	K	L
1	Items / Students	1	2	3	4	5	6	7	8	9	10	Student Total Score
2	Kid-A	1	1	1	1	1	1	1	1	0	1	9
3	Kid-B	1	1	1	1	1	1	1	0	1	0	8
4	Kid-C	1	1	1	1	1	1	0	1	0	0	7
5	Kid-D	1	1	1	1	1	0	1	0	1	0	7
6	Kid-E	1	1	1	1	1	1	0	1	0	0	7
7	Kid-F	1	1	1	0	1	0	0	0	0	0	4
8	Kid-G	1	1	0	1	0	1	0	0	0	0	4
9	Kid-H	1	0	1	0	1	0	0	0	0	0	3
10	Kid-I	0	1	1	0	0	0	0	0	0	0	2
11	Item total	8	8	8	6	7	5	3	3	2	1	

=sum(B2:K2)

=sum(B2:B10)

=L2-B2

Table 3: Computation of Total Score for Point-Biserial Correlation

		M	N	O	P	Q	R	S	T	U	V
1	Students	total-item1	total-item2	total-item3	total-item4	total-item5	total-item6	total-item7	total-item8	total-item9	total-item10
2	Kid-A	8	8	8	8	8	8	8	8	9	8
3	Kid-B	7	7	7	7	7	7	7	8	7	8
4	Kid-C	6	6	6	6	6	6	7	6	7	7
5	Kid-D	6	6	6	6	6	7	6	7	6	7
6	Kid-E	6	6	6	6	6	6	7	6	7	7
7	Kid-F	3	3	3	4	3	4	4	4	4	4
8	Kid-G	3	3	4	3	4	3	4	4	4	4
9	Kid-H	2	3	2	3	2	3	3	3	3	3
10	Kid-I	2	1	1	2	2	2	2	2	2	2

=CORREL(B2:B10, M2:M10)

Table 4: Computation of Total Score for Point-Biserial Correlation

		M	N	O	P	Q	R	S	T	U	V
		total-item1	total-item2	total-item3	total-item4	total-item5	total-item6	total-item7	total-item8	total-item9	total-item10
12	Point-Biserial	0.46	0.29	0.12	0.73	0.49	0.49	0.59	0.46	0.26	0.40
13	P-Value	0.89	0.89	0.89	0.67	0.78	0.56	0.33	0.33	0.22	0.11

=B11/9

Computing Point-Biserial Correlations Using SPSS

To compute the point-biserial correlation using SPSS, copy the sample data in Table 1 into an SPSS data window. Open a syntax window (File\New\Syntax), paste the following syntax and click Run.

```
RELIABILITY
/VARIABLES= item1 to item10  /FORMAT=NOLABELS  /SCALE(ALPHA)=ALL
/MODEL=ALPHA      /STATISTICS=SCALE    /SUMMARY=TOTAL.
```

The SPSS output window will show the following table. The column labeled *Corrected Item-Total Correlation* provides the point-biserial correlation.

Table 5: Item Point-Biserial Output from SPSS

Items	Scale Mean if Item Deleted	Scale Variance if Item Deleted	**Corrected Item-Total Correlation**	Cronbach's Alpha if Item Deleted
i1	4.7778	5.194	**0.457**	0.746
i2	4.7778	5.444	**0.286**	0.763
i3	4.7778	5.694	**0.122**	0.779
i4	5.0000	4.250	**0.728**	0.699
i5	4.8889	4.861	**0.486**	0.739
i6	5.1111	4.611	**0.491**	0.739
i7	5.3333	4.500	**0.589**	0.722
i8	5.3333	4.750	**0.459**	0.743
i9	5.4444	5.278	**0.260**	0.770
i10	5.5556	5.278	**0.399**	0.752

Corrected point-biserial correlation indicates that the item score is excluded from the total score before computing the correlation (as we did manually in Excel). This is a minor but important detail because inclusion of the item score in the total score can artificially inflate the point-biserial value (due to correlation of the item score with itself).

Interpretation of Results

Interpretation of point-biserial correlations

Referring to Tables 4 and 5, all items except for Item 3 – which has a point-biserial of 0.12 – show acceptable point-

biserial values. A low point-biserial implies that students who got the item incorrect also scored high on the test overall while students who got the item correct scored low on the test overall. Therefore, items with low point-biserial values need further examination. Something in the wording, presentation or content of such items may explain the low point-biserial correlation. However, even if nothing appears visibly faulty with such items, it is recommended that they be removed from scoring and future testing. When evaluating items it is helpful to use a minimum threshold value for the point-biserial correlation. A point-biserial value of at least 0.15 is recommended, though our experience has shown that "good" items have point-biserials above 0.25.

Interpretation of p-values

The p-value of an item provides the proportion of students that got the item correct, and is a proxy for item difficulty (or more precisely, item easiness). Refer to Table 4 and note that Item 10 has the lowest p-value (0.11). A brief examination of the data matrix explains why – only one student got that item correct. The highest p-value, 0.89, is associated with three items: 1, 2 and 3. Eight out of nine students got each of these three items correct. The higher the p-value, the easier the item. Low p-values indicate a difficult item. In general, tests are more reliable when the p-values are spread across the entire 0.0 to 1.0 range with a larger concentration toward the center, around 0.50. (Note: Better item difficulty statistics can be computed using psychometric models, but p-values give a reasonable estimate.)

The relationship between point-biserial correlations and p-values

Problematic items (items with a low point-biserial correlation) may show high p-values, but the high p-values should not be taken as indicative of item quality. Only the point-biserial should be used to judge item quality. Our sample data matrix contains two items that appear to have "conflicting" p-value and point-biserial statistics. One is Item 3, which has a low point-biserial (0.12) but a high p-value (0.89); the second is Item 10, which has a high point-biserial (0.40) but a low p-value (0.11).

Examination of the data matrix shows that Item 3 was answered incorrectly by Kid-G. Even though Kid-G did not correctly answer Item 3, she did correctly respond to Items 4 and 6, which are harder items (as indicated by their lower p-values). One explanation of how Kid-G could get the harder items correct but the easier item wrong is that she guessed on Items 4 and 6. Let us assume, however, that she did not guess and that she actually did answer Items 4 and 6 correctly. This would suggest that Item 3 measures something different from the rest of the test, at least for Kid-G, as she was unable to respond to this item correctly even though it is a relatively easy item. Although in this article we are dealing with a very small data matrix, in real life there may be a group of students like Kid-G. When faced with statistics such as we see here for Item 3 (high p-value, low point-biserial), it is recommended that the item be qualitatively reviewed for content and wording. Something in the wording of the item, its presentation or the content caused Kid-G to get it wrong and caused the low item point-biserial. In our little sample matrix, Item 3 is a problematic item because it does not *fit the model*, meaning that this item behaves differently from other items for Kid-G. The model says that Kid-G should have got that item right, but she got it wrong. Even if qualitative review of the item does not reveal any obvious reason for the low point-biserial, it is often advisable that this item be removed from future testing. Items that measure another content (also called multidimensionality) often show low point-biserials.

Now let us examine Item 10, which shows an opposite pattern. Item 10 has a high point-biserial (0.40) and a low p-value (0.11). This item was correctly answered by only one student (which explains the low p-value). However, the one student that got the item correct is also the "smartest" kid as measured on this test, which is why the point-biserial for the item is high. The low p-value and high point-biserial are perfectly acceptable statistics. The two numbers in this case tell us that Item 10 is a difficult item but not a problematic item.

Thus, there is no real relationship between the point-biserial correlation and p-value statistics. Problematic items will always show low point-biserial correlations, but the accompanying p-value may be low or high. **The point-biserial correlation should be used to assess item quality; p-values should be used assess item difficulty.**

Problematic items and test reliability

Let us briefly refer to one other statistic reported under the SPSS output in Table 2, labeled "Cronbach's Alpha if Item Deleted", which is an indicator of overall test reliability. Cronbach's Alpha ranges from 0 to 1, with a 0 indicating no test reliability and fractions approaching 1 indicating high test reliability. The SPSS output computes the reliability coefficient for the test excluding one item at a time. If the reliability increases when an item is deleted, that indicates that the item is problematic and reduces test reliability instead of increasing it. In our example, the test reliability is highest (0.779) when Item 3 is deleted. Remember that Item 3 was the most problematic item. This SPSS statistic emphasizes the point made earlier that removal of problematic items (misfitting items, multidimensional items, poorly written items) increases the overall test reliability.

Using point-biserial statistics to check multiple-choice keys

Another useful role of point-biserial correlations involves validating the multiple-choice scoring key. The key for all items on a test is communicated from item developers to form developers, then to test administrators and finally to test scorers. In this flow of information the scoring key may be incorrectly communicated or inadvertently mis-programmed into computerized scoring programs, which can cause serious errors in student results. One way to catch such errors is to run the point-biserial statistic on all items after the tests are scored. This is a quick, easy and reliable method for catching incorrect scoring keys. Items with incorrect keys will show point-biserials close to or below zero. As a rule of thumb, items with point-biserials below 0.10 should be examined for a possible incorrect key.