

# Interpreting your Test Scanning Services Report

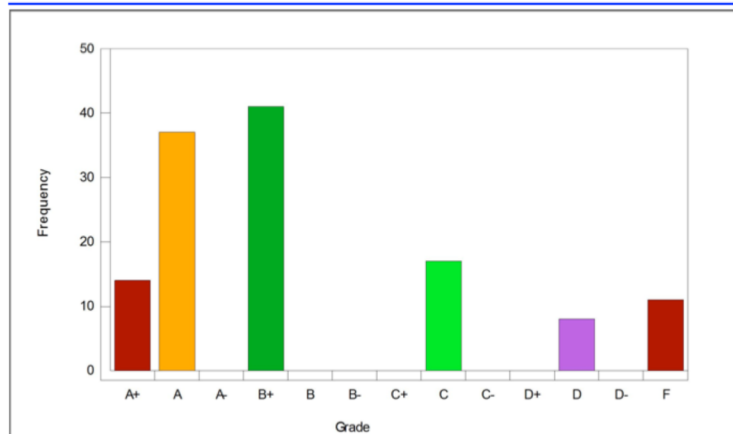
THE 'STUDENT RESPONSE REPORT' is a file that considers individual student's performance and displays it as a visual grid of student responses (rows) and questions (columns).

**Legend:** Correct: ■ Incorrect: ■ Unanswered or Other Incorrect Response: □

Students:	Question1	Question2	Question3	Question4	Question5	Question6	Question7	Question8	Total Score	Percent Score	Grade
<b>Answer Key:</b>	D	B	D	B	C	C	D	D	9.00	100.00	
Student name	<span style="color: green;">■</span>	<span style="color: green;">■</span>	<span style="color: green;">■</span>	<span style="color: green;">■</span>	<span style="color: green;">■</span>	<span style="color: green;">■</span>	<span style="color: green;">■</span>	<span style="color: green;">■</span>	8.00	88.89	A
Student name	<span style="color: green;">■</span>	<span style="color: green;">■</span>	<span style="color: green;">■</span>	<span style="color: green;">■</span>	<span style="color: green;">■</span>	<span style="color: green;">■</span>	<span style="color: green;">■</span>	<span style="color: green;">■</span>	8.00	88.89	A
Student name	<span style="color: green;">■</span>	<span style="color: green;">■</span>	<span style="color: green;">■</span>	<span style="color: green;">■</span>	<span style="color: green;">■</span>	<span style="color: green;">■</span>	<span style="color: green;">■</span>	<span style="color: green;">■</span>	6.00	66.67	C

THE 'CLASS FREQUENCY DISTRIBUTION REPORT' provides both a Table and a histogram of the distribution of grades for students who completed the test. A sample is shown here:

Overall			Mean Score: 77.26%	
Grade	Percent Score	Raw Score	Frequency	Percent
A+	90.00 - 100.00	8.10 - 9.00	14	10.94
A	85.00 - 89.99	7.65 - 8.09	37	28.91
A-	80.00 - 84.99	7.20 - 7.64	0	0.00
B+	77.00 - 79.99	6.93 - 7.19	41	32.03
B	73.00 - 76.99	6.57 - 6.92	0	0.00
B-	70.00 - 72.99	6.30 - 6.56	0	0.00
C+	67.00 - 69.99	6.03 - 6.29	0	0.00
C	63.00 - 66.99	5.67 - 6.02	17	13.28
C-	60.00 - 62.99	5.40 - 5.66	0	0.00
D+	57.00 - 59.99	5.13 - 5.39	0	0.00
D	53.00 - 56.99	4.77 - 5.12	8	6.25
D-	50.00 - 52.99	4.50 - 4.76	0	0.00
F	0.00 - 49.99	0.00 - 4.49	11	8.59



THE 'TEST STATISTICS REPORT' provides a synopsis of the test attributes and some important statistics. A sample is shown here to the right.

Overall	
<b>Score Data</b>	
Number of Graded Items	44
Total Points Possible	44
Maximum Score	41
Minimum Score	18
<b>Statistics</b>	
Mean Score	31.37
Mean Percent Score	71.30
Benchmark Score	-
Range of Scores	23.00
Standard Deviation	5.89
Variance	34.71
<b>Percentiles</b>	
Percentile (25)	27.00
Median Score	32.00
Percentile (75)	36.00
Inter Quartile Range	9.00
<b>Confidence Intervals</b>	
1%	28.68
5%	29.36
95%	33.38
99%	34.06
<b>Test Reliability</b>	
Kuder-Richardson Formula 20	0.82
Kuder-Richardson Formula 21	0.76
Coefficient (Cronbach) Alpha	0.81

The Test reliability indicators are measures of how well the questions individually and collectively measure the same thing (hopefully content mastery). (more about validity and reliability can be found at the end of the report descriptions).

**Kuder-Richardson Formulas:** Are formulae for testing reliability as a measure of internal consistency. Higher values indicate a stronger relationship between assessment items.

- Values range between 0.0 (unreliable) and 1.0 (completely reliable).
- Scores above .8 are very reliable
- Scores below .5 are not considered reliable

**Coefficient (Cronbach) Alpha:** A coefficient of reliability reflecting how well the assessment items measure internal consistency for the test.

- Score  $\geq$  .9 Excellent
- .9 > score  $\geq$  .8 Good
- .8 > score  $\geq$  .7 Acceptable
- .7 > score  $\geq$  .6 Questionable
- .6 > score  $\geq$  .5 Poor
- .5 > score < .5 Unacceptable

These reliability scores are generated by determining how students in the top quartile perform relative to the lowest quartile of students on all of the individual questions. For example you would expect that for each individual question more students in the top quartile will get the question correct more frequently than students in the lowest quartile. A generally reliable test might have a 'lower than acceptable' reliability due to a few flawed questions. The remaining reports can help identify these.

The remaining reports are aimed at identifying likely flawed questions. To read these reports you need to know that the ‘distractors’ is the name given to all of the incorrect choices provided on the test. Effective distractors are chosen by some students and are most often chosen by students who overall do not do well on the test.

**THE ‘CONDENSED TEST REPORT’** Gives an overall reliability score for the test and then a question analysis. Each row provides information on a specified question and displays the frequency that choice was selected, as well as the total % correct responses.

**Legend:** Distractors Chosen More than Correct Answer: ■

<b>Total Possible Points:</b>	9.00	<b>Median Score:</b>	4.00	<b>Maximum Score:</b>	9.00
<b>Total Students:</b>	50	<b>Mean Score:</b>	4.30	<b>Minimum Score:</b>	1.00
<b>Standard Deviation:</b>	1.83	<b>Reliability Coefficient (KR20):</b>	0.48	<b>Range of Scores:</b>	8.00

No.	Question	Correct Answer	Response Frequencies					Non Distractor	Correct Group Responses			Point Biserial
			A	B	C	D	E		Total %	Upper 27%	Lower 27%	
1	Question1	C	8.00	18.00	50.00	2.00	22.00		50.00	78.57	28.57	0.47
2	Question2	C	42.00	2.00	24.00	18.00	14.00		24.00	42.86	7.14	0.37

Each row also provides the percentage of students in both the top and bottom quartiles who selected the correct choice. Flawed questions have more lowest quartile-students selecting it than upper quartile students. (lower and upper quartile students are determined by their score on the test). The point-biserial is a statistic that quantifies how well student performance on this question correlates with their overall test performance.

The higher the biserial value, the greater the question’s ability to discriminate between those students who generally know the test material and those who do not.

0.40 or higher very good items

0.30 to 0.39 good items

0.20 to 0.29 fairly good items (0.25 is one recommended cut off for useful questions, but higher is better)

0.19 or less poor items

In general the overall reliability of the test is improved if poor questions are removed.

But Mehrens and Lehmann (1973) offer three cautions about using the results of item analysis:

1) Item analysis data are not synonymous with item validity. An external criterion is required to accurately judge the validity of test items. By using the internal criterion of total test score, item analyses reflect internal consistency of items rather than validity.

2) The discrimination index is not always a measure of item quality. There are a variety of reasons why an item may have low discrimination power:

(A) extremely difficult or easy items will have low ability to discriminate, but such items are often needed to adequately sample course content and objectives.

(B) an item may show low discrimination if the test measures many content areas and cognitive skills. For example, if the majority of the test measures "knowledge of facts," then an item assessing "ability to apply principles" may have a low correlation with total test score, yet both types of items are needed to measure attainment of course objectives.

(C) Item analysis data are tentative. Such data are influenced by the type and number of students being tested, instructional procedures employed, and chance errors. If repeated use of items is possible, statistics should be recorded for each administration of each item.

Rinehart and Winston, 333-334.

**THE 'TEST ITEM STATISTIC REPORT'** is way of summarizing the performance of each question without looking at the results for each separate choice; it lumps all incorrect answers together. This report quickly allows you to see the percentage of the class that got the question correct and the point biserial value of the question.

Question		Summary Statistics				Reliability	
Question	Points	Graded	Correct	Incorrect	No Response	Point Biserial	Percent Correct
Question1	1.00	35	31	4	0	0.10	88.57
Question2	1.00	35	31	4	0	0.31	88.57
Question3	1.00	35	28	7	0	0.54	80.00
Question4	1.00	35	27	8	0	0.31	77.14
Question5	1.00	35	34	1	0	0.27	97.14
Question6	1.00	35	31	4	0	0.07	88.57

**THE 'DETAILED ITEM ANALYSIS REPORT'** is yet another way to look at the consistency of questions, but this report let's you look at the consistency of both correct and incorrect choices. The correct choice (indicated with an asterisk) should have a relative high positive value and distractors should have a low value.

Question1

Response	Frequency	Percent	Point Biserial
A	4	8.00	0.03
B	9	18.00	-0.25
*C	25	50.00	0.47
D	1	2.00	-0.02
E	11	22.00	-0.35
<b>Total</b>	<b>50</b>	<b>100.00</b>	

Lastly the **'CONDENSED ITEM ANALYSIS REPORT'** gives you a quick visual look at how often each option was selected.

Legend: Correct: ■ Incorrect: ■ Distractors Chosen More than Correct Answer: ■

Response	Frequency	Percent
A	3	8.57
B	3	8.57
C	15	42.86
*D	13	37.14
E	1	2.86
Missing	0	0.00

## INTERPRETING TEST REPORTS

**Effective testing:** An effective test is both accurate and consistent when measuring how well students have mastered course material. This means:

1. The test is measuring content the instructor has identified as being covered on the assessment.
2. The assessment is built so students who have mastered the identified content are more likely to achieve higher scores than students who have not mastered the identified content.
3. An inaccurate and inconsistent test does a poor job of assessing student mastery of content, often through unclear assessment items, returning results where students who have not mastered the content might achieve scores equal to or even surpassing students who have greater mastery of the content.

How can test reports help identify poor tests or poor test items (individual questions)?

**Internal consistency:** Measures the correlation of all assessment items in a test, presented as a value.

1. A scaled score used to identify whether a pool of items consistently measures the mastery of course content.
2. The scale offers a continuum between high internal consistency or low internal consistency.
3. The higher the reported internal consistency, the greater correlation with how well individual test items measure the same content.

**Reliability:** Represents how consistently an assessment returns the same values.

1. Students taking the same test multiple times would be expected to achieve similar scores with each attempt unless additional information was presented or known material is reinforced between assessment events.
2. An unreliable test is more likely to return random scores with each attempt.

**Validity:** While validity is not explicitly measured in our reports, it should be pointed out as a meaningful concern when designing an assessment.

1. A valid test is one that accurately measures the required material.
2. Another way to view this is that a valid test would not contain items that would surprise a student who has a strong mastery of the content identified as being covered in the test.
3. Reliability and validity are tied closely together, but are not synonymous.
4. While our reports are not able to quantify a value for validity, there is a strong connection between the two.
5. A test showing a high mark for validity would likely return a high mark for reliability. But while a high mark for reliability MIGHT correlate to a high mark for validity, it DOES NOT assure validity. The data and scores provided in the reports either focus on the overall assessment (test) or the individual assessment items (questions).

While the reports provide a wealth of data the primary combination of measurements is the Coefficient (Cronbach's) Alpha and Kuder Richardson 20, both measuring the internal consistency of the assessment, and the point biserial, measuring the reliability of each assessment item. Generally acceptable values for the point-biserial statistic is given in the **'CONDENSED TEST REPORT'**, and values for the others are found within the **'TEST STATISTICS REPORT'**

Haladyna, T. M. (1999). *Developing and validating multiple-choice test items* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Questions about interpreting reports? For now ask Clare Hasenkampf [hasenkampf@utsc.utoronto.ca](mailto:hasenkampf@utsc.utoronto.ca)