Quick start guide for using the test level report produced by Integrity

http://integrity.castlerockresearch.com
Castle Rock Research Corp.
November 17, 2005

The test level report is designed to provide information on the overall performance of your test. The test level report has five sections of information:

1) The Summary Section, which provides a written/text description of some key aspects of the test performance
2) The Table of Statistics Summary, which provides statistical summary information of the test performance
3) The Histogram of Test Total Score, which provides a graphical representation of the distribution of total test scores (e.g., how many students obtained different scores on the test)
4) The Frequency Distribution of Test Scores, which provides a numeric summary of the number of students who obtained different scores on the test
5) The Cumulative Percent of Total Test Score graph, which provides a graphical summary of the percentage of students who obtained different raw scores on the test.

The statements in the Summary Section are designed to flag the potentially problematic aspects of the test to you, the user, as well as to indicate if the test is performing well. The statements should guide you to different parts of the reports in order to investigate why a problem is occurring and give suggestions for corrective action. The Table of Test Statistics is designed to provide detailed statistical information regarding the performance of the test. An example of these two sections is presented in Figure 1.

Figure 1. An example of the Summary and Table of Statistics summary sections of a test-level report produced by Integrity.

Summary

→ The test mean is very low. Consider examining the items on the test to determine whether they are too difficult for the examinees being tested.

→ The KR-20 for this test indicates low test reliability. The reliability of the test is related to factors such as: 1) low number of test items, 2) small number of examinees, 3) many items that are too difficult to too easy, 4) many items that have low discrimination, 5) the items on the test are not measuring one dominant trait, 6) students do not have enough time to finish all the items on the test. Consider investigating the above factors in order to increase test reliability.

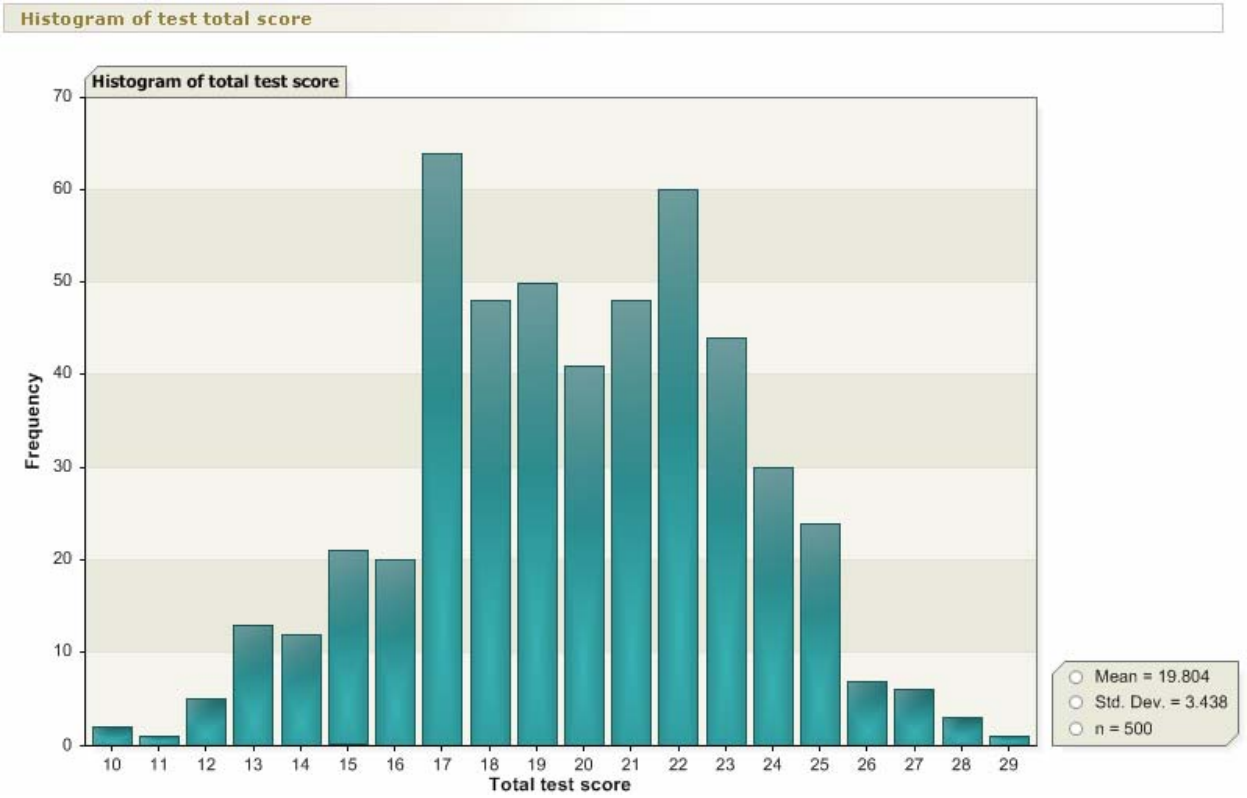
→ The skewness and kurtosis statistics indicate that the distribution of scores is relatively normally distributed.

Table of statistics summary

Number of examinees = 500	**Standard error of mean** = 0.154
Number of items on test = 60	**Standard error of measurement** = 3.022
Mean = 19.804	**KR-20 reliability** = 0.227
Median = 20.000	**Spearman-Brown split-half reliability coefficient** = 0.223
Mode = 17.000	**Spearman-Brown prophecy reliability formula** = 0.365
Standard deviation = 3.438	**Guttman split-half reliability coefficient** = 0.220
Variance = 11.817	**Skewness (total score)** = -0.122
Maximum score = 29	**Kurtosis (total score)** = -0.341
Minimum score = 10	

The first summary statement refers to the mean of the test and is directing you to look at the mean value in the Table of Statistics summary. The mean value in the table is 19.804 out of 60 items, which indicates that the percentage test mean is 33%. The mean is a measure of central tendency, so it describes what test score was in the middle range out of all the possible test scores (e.g., out of all possible scores from 0 to 60, most students scored about 20 out of 60). The Histogram of Test Total Score (see Figure 2) shows the distribution of student scores presented visually. You can see that most of the scores on the test are near the value of 22, with no students scoring over 29 out of 60.

Figure 2. An example of a Histogram of total test score produced by Integrity.



The Frequency Distribution of Test Scores should show the number of students who achieved each possible test score (in the example shown, 12 out of 500 students obtained a score of “14” on the test, 64 students obtained a score of “17” on the test, 60 students obtained a score of “22” on the test, etc.). The histogram is a visual representation of the frequency distribution of scores. An example of the Frequency Distribution of Test Scores is presented in Figure 3.

Figure 3. An example of a Frequency distribution of test scores produced by Integrity.

Frequency distribution of test scores

Raw score	Frequency	Percent	Cumulative percent
10	2	0.400	0.400
11	1	0.200	0.600
12	5	1.000	1.600
13	13	2.600	4.200
14	12	2.400	6.600
15	21	4.200	10.800
16	20	4.000	14.800
17	64	12.800	27.600
18	48	9.600	37.200
19	50	10.000	47.200
20	41	8.200	55.400
21	48	9.600	65.000
22	60	12.000	77.000
23	44	8.800	85.800
24	30	6.000	91.800
25	24	4.800	96.600
26	7	1.400	98.000
27	6	1.200	99.200
28	3	0.600	99.800
29	1	0.200	100.000

There are a number of possibilities as to why the mean of this example test is so low. It could be that the key that was submitted to score the test was incorrect (e.g., the key that was submitted was for another test). If the key is incorrect, most of the student responses to items would be scored incorrect even though they may have been correct. You should double-check the key file you submitted to ensure that it was the correct key. Another option is that the key is correct and that the material on the test was far too difficult for the group of students being tested. In this case, you may wish to examine the item level reports that Integrity produces in order to identify which items students found the most difficult on the test. You should then review the content of each item on the test in order to determine why the group of students found most of the items too difficult (e.g., perhaps the group of students being tested are at a remedial level but were presented with advanced material beyond their ability).

An example of a summary statement that would direct you to investigate the performance of your test further is:

> "The distribution of scores is highly negatively skewed and therefore not normally distributed. This indicates that there is a higher density of examinees obtaining higher scores than moderate or low scores."

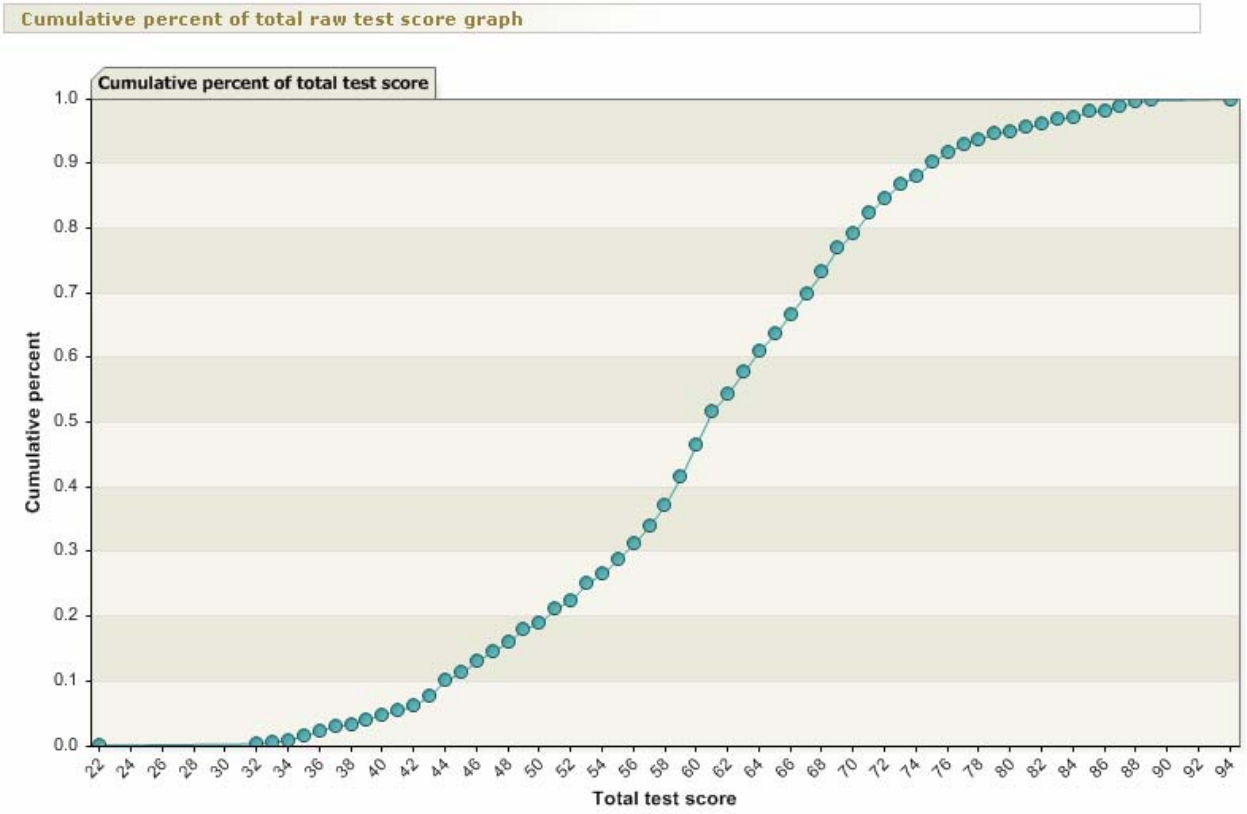
This statement is based on the "skewness" statistic in the Table of Statistics summary. Skewness is a measure of the symmetry/asymmetry of distribution of test scores. It provides information on how much a distribution is "pushed" to one side or another. In this case, the distribution is highly negatively skewed, which means that there are many more students obtaining higher scores on the test than students obtaining middle or low scores. You should see this pattern of the distribution of scores being pushed to one end in the Histogram of Test Total Score graph.

There are a number of reasons why the test results can be skewed. One reason could be that the class of students being tested is a high-performing group, and therefore the majority of the students in the class received high scores. In this case, there may be nothing "wrong" with your assessment; in fact, it could mean that you are doing an outstanding job of instruction. Another potential explanation is that the test is not challenging enough for the students who took the assessment. It could be the case that in an effort to not make the assessment too difficult, it was made too easy, and therefore the majority of students got most of the questions on the test correct. The only way to determine the actual cause of the skewness is to investigate the issue further. You should take a look at the item level reports and see which items students found simple and which items students found challenging and hypothesize as to the cause (e.g., "Almost all students got item 5 correct. The content of that item was covered very thoroughly in class so that could be why most students got it right"). This investigative process whereby you examine the content of each item based on the item statistics may shed light on what test level patterns are occurring. As an aside, generally you would expect assessment results that are not skewed but rather resemble a bell shaped curve (called a normal distribution). This is because it is not usually expected that all the scores should be bunched up at one end of the score distribution. Rather, it is generally expected that scores be distributed around the mean fairly evenly. The intended purpose and design of the assessment are what should be supported by the report information in Integrity. If the intended purpose and design of the assessment are not supported (e.g., the exam is much more difficult than was intended), then investigating why and taking steps to correct the problems is the appropriate course of action.

The Cumulative Percent of Total Test Score graph, an example of which is shown in Figure 4, provides information on the cumulative percent of scores that were obtained by all students who wrote the assessment. The information contained in cumulative percentage column of the Frequency Distribution of Test Scores is plotted in the Cumulative Percent of Total Test Scores graph. This graph is usually expected to be an "S-type" curve where the percentage of low total test scores rises and plateaus in the high score range. This pattern indicates that fewer students received low scores, the

majority of students received middle scores, and few students received high scores. Large plateaus in the graph may indicate problems with the test in that there are large gaps in the percentage of students obtaining scores in a particular test score range. For example, if no students obtain scores between 50 and 65 (out of 100 items), causing plateaus to be seen in the graph, this is grounds for concern as it would generally be expected a large number of students would achieve scores in the middle area of the score distribution.

Figure 4. An example of a Cumulative Percent of Total Test Score graph produced by Integrity.



What else should I look for in the test level reports?

By looking through some of the information contained in the Table of Statistics Summary, you can gain insight into the performance of your test. For example, take a look at the “Maximum score” and “Minimum score” sections. This tells you what the highest score achieved on the test and the lowest score achieved on the test were. It may be the case that the highest score is 94 out of 100. This may prompt you to look into why no students, not even your top students, achieved 100 out of 100 on the test. By going to the Examinee Results report of Integrity, you can look through the scores for all the students in your class and see whether the top students are performing as you would have expected. If your top students are not obtaining the scores you would have expected, it may be an opportunity to find out why (e.g., perhaps some items were ambiguous and caused confusion for the top students).

The Frequency Distribution and Histogram of Total Test Scores are useful in examining the test scores that students achieved. The histogram is a graphical representation of the frequency distribution results. If you have set a score of 50 out of 100 as the “pass mark” for a test, it may be useful to see how many students were on the borderline of passing the test (e.g., students who obtained scores of 45 to 49). It may be the case that there are a number of students in this borderline passing range that could be moved over the passing threshold with some extra attention on certain topics. You can then go into the Examinee Results report to identify students that received scores in this borderline range and develop strategies to improve their performance.

The standard deviation (Std. Dev.) provides information on how “spread out” test scores are. For example, if the mean of a 100 item test is 65 and the standard deviation is 2, this indicates that very few students obtained scores that were much different than 65. If, on the other hand, the standard deviation of the test was 20, this would indicate that the scores students obtained on the test varied a great deal from 65. The standard deviation is not really a statistic that lends itself well to answer the question “What is an acceptable standard deviation for my test?” You should be watchful of very large standard deviations and very small standard deviations because these indicate that the scores on the test are respectively too disperse or hardly dispersed at all. By looking at the Histogram of Test Total Score, you will be able to visualize how the scores on the test are dispersed. If there is little variation in test scores (i.e., very small standard deviation), this could indicate problems with the spread of the difficulty of items that compose the test (e.g., perhaps all the items on the test are very similar in difficulty, resulting in little variation in test scores). If there is a great deal of variation in test scores (i.e., very small standard deviation), this could indicate the opposite problem with the spread of items on the test (e.g., perhaps the items on the test have very diverse difficulties, resulting in a great deal of variation in test scores). The item level reports would provide useful information in terms of the difficulty spread of items.

The KR-20 reliability coefficient, Spearman-Brown split-half reliability coefficient, and Guttman split-half reliability coefficient provide information on the reliability of the assessment. Reliability in this context refers to how consistently the test measures what it is supposed to measure. For example, imagine if there were a device that could erase memories. A student took an assessment, then, using this device, had his memory of having taken the test erased. A week later he took the test again. The correlation between the scores the student obtained on the test on these two occasions would be an indicator of the reliability of the assessment. A high similarity (correlation) between the student’s first score and his second score would indicate high reliability, because the assessment yielded similar measurements of the student’s knowledge or ability. As there is no memory-erasing devise in existence (not to mention it would be unethical to use such a device to determine test reliability), statistical methods are used to estimate the reliability of the test in different ways. For example, the Spearman-Brown split-half reliability coefficient uses the average correlation between random split-halves of the test.

The reliability of an assessment is related to factors such as:

- The number of assessment items – Reliability is related to test length in that tests with more items (measurements) tend to have greater reliability. The Spearman-Brown prophecy reliability formula provides a Spearman-Brown reliability coefficient if the test length were doubled.
- The number of examinees who took the assessment – Greater numbers of students can increase the reliability coefficient of the test because the group of students tested would be more representative of the entire population of students.
- The difficulty of the items that compose the test – Having many items at the extreme ends of the difficulty continuum (e.g., very difficult and very simple items) will decrease test reliability.
- The discrimination of items that compose the test – This is related to test reliability in that if many items on the test have low discrimination, reliability will be decreased. This is discussed in more detail in the item-level guide.
- Items that compose the test not measuring one dominant construct – This is related to test reliability in that if the test is composed of many sub-domains split-halves of the items may not yield high correlations.
- Students not having enough time to finish all the items on the test – This is related to reliability in that some items may not have the opportunity to provide measurements regarding the students' knowledge and ability.

If the reliability indexes for your test are low (e.g., less than 0.70), consider examining these factors to try and improve the tests reliability.

