

How Integrity can help evaluate and improve the quality of assessments

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

Why evaluate and improve the quality of assessments?

In classrooms all over the world assessments are used to evaluate how well students have learned, retained, and mastered course material. How well an assessment accomplishes the goal of estimating a student's knowledge or ability is directly related to the quality of the assessment. If a test is composed of questions that do not measure student knowledge or ability well, then the instructor will not obtain an accurate representation of how well the students are learning and retaining the course material. As a result, the instructor may inadvertently be providing students with grades that do not reflect their "true" knowledge of the course material.

Diagnosing, evaluating, and improving the quality of measurement instruments is important in any field. For example, the importance of conducting routine diagnostics and servicing on the electronic measurement instruments on an airplane is self-evident. Airlines must ensure that measurements of altitude, airspeed, and engine temperature are as accurate as possible so that the safety of the passengers and crew is guaranteed. Auto mechanics routinely evaluate and calibrate their diagnostic computers to ensure that they are providing accurate measurements of the performance of the vehicles. There are benefits to evaluating and improving the performance of all measurement instruments and there are negative consequences to not doing so.

In the domain of educational assessment, scientific measurements of how well tests and test items perform involve the use of statistics. These statistics were designed by educational and psychological measurement experts, some of whom are referred to as psychometricians. Accurate measurements of student achievement provide important feedback to students about how they can improve their learning, as well as to educators in terms of how they can improve their instruction.

How Integrity can help

Integrity is designed to provide useful information to you whether you want to learn more about how to improve your assessments or you are an experienced expert in the field of educational and psychological measurement. Integrity contains a detailed glossary of terms: by clicking on any term, you can view its definition and find out more information on how to interpret the associated statistics or graphs.

The reports produced by Integrity offer detailed yet accessible information regarding the performance of individual test items and of the test overall. In addition, Integrity offers

breakdowns at the student (examinee) level, subscale level (e.g., curriculum content areas), group level (e.g., male versus female), writing center level (e.g., class 1 versus class 2), and more.

Examples of reports produced by Integrity

The executive summary report

The three sections of the executive summary report are designed to provide a quick overview of the key findings of the assessment analysis:

- 1) The **Summary section** provides statements that summarize key aspects of the test and flag potential problems with test items or with the test as a whole. An example of the **Summary section** of the executive summary is shown below. In this example, item 2 has been flagged as potentially being too difficult an item. Item 2 has also been flagged as having a negative discrimination value (CPBR) and, therefore, as possibly being mis-keyed. By clicking on the item number, you will be taken to a detailed item level report where these statements are expanded upon and other characteristics of item 2's performance are presented.

The summary statements can also be used to obtain more information. For example, one of the summary statements is "The KR-20 for this test indicates moderate to high test reliability." If you are interested in finding out more about what KR-20 test reliability is and how it relates to your test performance, go to the next section of the executive summary, the **Table of test statistics**, and click on the term. You will be presented with the definition of the term, retrieved from the glossary.

- 2) The **Table of test statistics** provides statistical information such as the maximum and minimum scores achieved on the test, the mean of the test, and other detailed information. Information such as the "Maximum score" is useful because it tells whether or not the top students obtained the highest possible score on the assessment. If no students did, you may want to investigate further by, for example, checking the results of these top students to find out if they all got the same few questions wrong. This might lead you to check the wording and clarity of those few questions or to reconsider whether those concepts were taught thoroughly enough. In similar manner, the "Minimum score" is useful in telling how struggling students performed on the assessment. All terms in the **Table of test statistics** are clickable and thoroughly defined through the glossary.
- 3) The **Histogram of total test scores** provides a visual representation of the number of students who achieved each test score. For example, it shows the number of students who obtained a score of 35 out of 55, 36 out of 55, and so on. This graph is useful in evaluating the distribution of test scores. For example, if the histogram shows that many students received very high scores on the test,

it may be that either the test was too easy or that the students mastered the test material at a high level. In any case, further investigation would be appropriate. For example, you might review the content of the test items and refer to the item statistics report to evaluate the difficulty of the items – “Are the items less difficult than I thought they were?”.

An example of a summary section within the executive summary produced by Integrity.

Summary	
Summary statements	Applicable items
→ This item may be too difficult. Item difficulty affects item discrimination in that items of high and low difficulty may have lower discrimination statistics. Consider reviewing the content of the item to determine if it should be made less difficult.	2
→ This item has a negative CPBR, which is statistically very problematic. Examinees of low ability may have a greater probability of answering the item correctly, than do examinees of high ability. This item may be keyed incorrectly. If the item key is correct, review the content of the item carefully and consider deleting this item from the test.	2
→ This item has low discrimination. Examinees of low ability should have a much lower probability of answering an item correctly than do examinees of high ability. Low discrimination statistics suggest that this may not be what is occurring. Consider reviewing and revising the content of this item to see if ambiguity in the item content can be limited. Also, consider that item difficulty affects item discrimination in that items of high and low difficulty may have lower discrimination statistics.	3 12 13 14 15 16 17 20 22 23 24 25 36 39 40 41 42 43 45 47 48 49 50 51 52 53 55 56 57 60 61 63 64 65 66 68 70 80 81 83 84 85 86 87 88 90 91 92 94 96 97 98 99
→ This item may be too easy. Item difficulty affects item discrimination in that items of high and low difficulty may have lower discrimination statistics. Consider reviewing the content of the item to determine if it should be made more difficult.	17 42 83
→ The KR-20 for this test indicates moderate to high test reliability.	
→ 3 pairs of examinees have been identified by the collusion detection analysis.	

Item-level statistical reports

By clicking on an item listed in the Executive Summary report, you are taken to a detailed item-level statistical report. It begins with an information bar that shows the item number, keyed correct answer to the item, difficulty level of the item, and item discrimination. This information bar is followed by item-specific summary statements that describe characteristics of the items and direct you to other sections of the item level report. An example of the information bar and summary statements is shown below.

An example of summary statements and the information bar from an item-level report produced by Integrity.

Item 51 - Key = B , Difficulty = 0.598 , Discrimination (CPBR) = 0.195
→ This item is of moderate difficulty.
→ This item has low discrimination. Examinees of low ability should have a much lower probability of answering an item correctly than do examinees of high ability. Low discrimination statistics suggest that this may not be what is occurring. Consider reviewing and revising the content of this item to see if ambiguity in the item content can be limited. Also, consider that item difficulty affects item discrimination in that items of high and low difficulty may have lower discrimination statistics.
→ Compared with other incorrect alternatives, few examinees selected incorrect alternative D for this item. Consider reviewing the content of this alternative to determine if it can be revised to attract more examinees.

The first summary statement for this item states that the item is of moderate difficulty. According to the information bar, the difficulty for the item is 0.598, which means that 59.8% of students who took the test selected the correct answer (in this case “B”) to this item.

The second statement has to do with the discrimination of the item. In general terms, item discrimination refers to how well items differentiate between students of different ability levels. The higher the discrimination statistics, the better the item differentiates between the top-, middle-, and low-performing students. For more information on the concept of item discrimination, click on the term within Integrity. The summary statement offers some direction in terms of what to look for in your item to try to evaluate and improve the performance of the item.

The third summary statement for this example relates to the performance of one of the item alternatives: option D. The summary statement describes a potential problem with the item in that very few students selected option D. Upon reviewing this alternative, you may find that it is so obviously incorrect that no students are choosing it.

By reviewing the content of the item-specific summary statements, you can gain a better understanding of the strengths and weaknesses of items and how to improve items.