Quick start guide for using Integrity item specific reports to improve assessment items

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

The item specific reports contain a wealth of information regarding the performance of individual items. As items are building blocks of the assessment, evaluation of the performance of each item is necessary in order to improve the overall assessment.

The concepts of item difficulty and item discrimination

Two concepts that are important when examining the performance of test items are item difficulty and item discrimination. Item difficulty in this context refers to the proportion of students who selected the correct response. For example, an item difficulty value of 0.645 indicates that 64.5% of students selected the correct response.

Item discrimination refers to how well items differentiate (i.e., discriminate) between students of different ability levels. The indexes that are used to evaluate the discrimination performance of items are correlations between the correct/incorrect examinee responses to items and the total test scores of examinees. In other words, the correlations are used to determine how much of a relationship exists between whether students get an item correct or incorrect and the students' total test scores. For example, one would expect that the highest-achieving students in a class would generally have high test scores. In order to obtain high test scores, these students would need to get many items correct on the test. In similar manner, one would expect that the lowest-achieving students in a class would generally have low test scores. These low test scores occur because these students get many items incorrect. Students in the middle range of performance would be expected to have middle-of-the-road test scores, obtaining approximately half of the items on the test correct. It follows, then, that high-achieving students (students with high total test scores) would be expected to have a higher likelihood of selecting the correct response to items than would middle-achieving or low-achieving students. In similar manner, middle achieving students (students with moderate total test scores) would have a higher likelihood of selecting the correct response to an item than would low-achieving students. This "pyramiding" response pattern is what the discrimination statistics represent and what one should look for in a discriminating item (higher discrimination statistics equal higher discrimination).

The academic research literature varies in terms of what discrimination indexes are recommended and the acceptable values for those indexes. Integrity uses the corrected point-biserial correlation (CPBR) index as the main indicator of discrimination and we have chosen values of 0.225 and below to indicate low discrimination, values of between 0.225 and 0.350 to indicate moderate discrimination, and values of 0.350 and greater to indicate high item discrimination.

What is considered “acceptable” test difficulty is dependant on what the assessment is designed to do. For example, if the assessment is designed to be challenging for the majority of students and the instructor expects that the average score on the test should be around 65%, then the assessment should be composed of items with a range of difficulties with the majority of items in the middle range of difficulty. Generally tests are designed to be composed of items with a range of difficulties with a few relatively difficult items, a few relatively easy items, and most items in the middle range. Integrity uses difficulty values of 0.35 and below to signify “Highly” difficulty items, difficulty values between 0.35 and 0.75 to signify “Moderately” difficult items, and difficulty values of 0.75 and above to signify “Low” difficulty items.

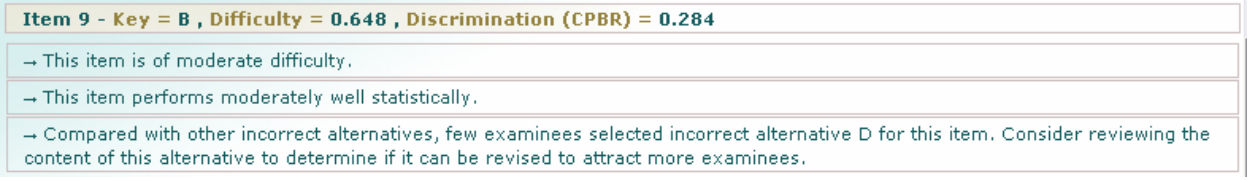
Item specific statistical report

The item summary statements, which appear in the executive summary and the item summary pages, provide written summaries of how items on the test perform. The summary statements are designed to flag potentially problematic items and to identify features of items. Next to the summary statements is a list of the items to which the statements apply. By clicking on any of the items, you can go into the detailed item reports for that item and find out more about its performance. The item-specific statistical report provides specific item information in four sections:

1) item-specific summary statements
2) correlation coefficients (discrimination statistics)
3) group breakdowns of how many students selected each item alternative
4) item performance plot

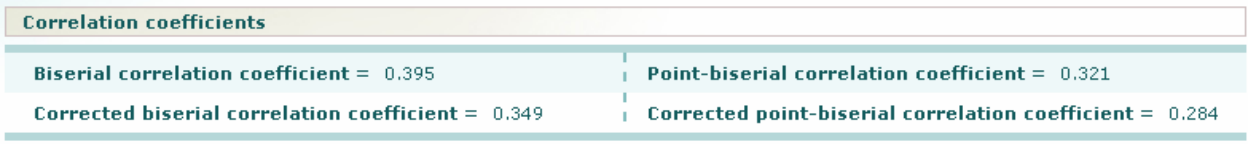
The item-specific summary statements provide information regarding the performance of a specific item. The summary statements are designed to describe characteristics of the items and direct you to the other sections of the item-specific report. Above the summary statements is an information bar that shows the item number, keyed correct answer to the item, difficulty level, and item discrimination. An example of summary statements and the information bar is shown in Figure 1.

Figure 1. An example of summary statements and the information bar from an item specific report produced by Integrity.



The correlation coefficients are all different measures of discrimination. The corrected point-biserial correlation coefficient is the measure of discrimination used throughout Integrity as the main measure of discrimination. An example of the correlation coefficients section of the detailed item specific report are shown in Figure 2.

Figure 2. An example of the correlation coefficients section of an item specific report produced by Integrity.



The group breakdowns table provides information on what proportion of students selected each item alternative (e.g., A, B, C, D). The rows of the table show the proportion of students who did not select an answer for the item and the proportion of students who selected each item alternative. The correct answer is bolded. There are five columns in the table:

i. Total – This column lists the proportion of all students who took the test who selected each item alternative.
ii. Top – This column lists the average proportion of the top quartile (top 25%) of examinees (according to their total test scores) who selected each item alternative.
iii. Mid. – This column lists the average proportion of the middle range students (between the 75% and 25% percentiles according to their total test scores) who selected each item alternative.
iv. Low – This column lists the average proportion of the lower quartile (bottom 25%) of students (according to their total test scores) who selected each item alternative.
v. TTS – This column lists the average total test scores for all students who selected each item alternative.

This group breakdown information is useful in investigating how many students selected each item alternative. For example, if we are examining an item of moderate difficulty, such as the one reported in Figure 1, we can quickly identify patterns of responding that may shed light on how to improve the item. The group breakdown table in Figure 3 shows that 64.8% (0.648 in the Total column) of all students selected the correct response (B) to the item. No students did not fill in an answer to the item,15.4% of all students selected A, 18.8% of all students selected C, and 1% of all students selected D. Immediately, we can see that option D drew hardly any students to it. This item response option should be examined to determine whether it can be revised to draw more students to it. By moving across the columns, we can see how the top, middle, and low students in the class responded to the item. For example, for the correct response option (B), 82.2% of the top students selected it, 67.5% of the middle students selected it, and 43.4% of the low students selected it. We can see that the top group of students in the class had the highest proportion of students selecting the correct answer, followed by the middle group of students, followed by the low group of students. This “pyramid” pattern is what one would expect in a discriminating item: the item differentiates between students of different performance levels. For the incorrect alternatives, we should see the opposite pyramid pattern: low-performance students should select the incorrect alternatives more than the middle students, and the middle students should select the incorrect alternatives more than the top students. If we do not

find this pattern (e.g., one of the incorrect alternatives draws more top students than middle or low students), this is cause for concern as top students may be "second guessing" themselves as to what the correct answer is. Finally, the TTS column shows what the average total test scores were for students that selected each item alternative. We would expect that the average total test score for the students that selected the correct response option should be higher than for students that selected any of the incorrect response options. In this example, we find this is the case: the average total test score for students that selected the correct alternative was 63.825 with the average total test score for each of the incorrect alternatives much less.

Figure 3. An example of a group breakdowns table for an item produced by Integrity.

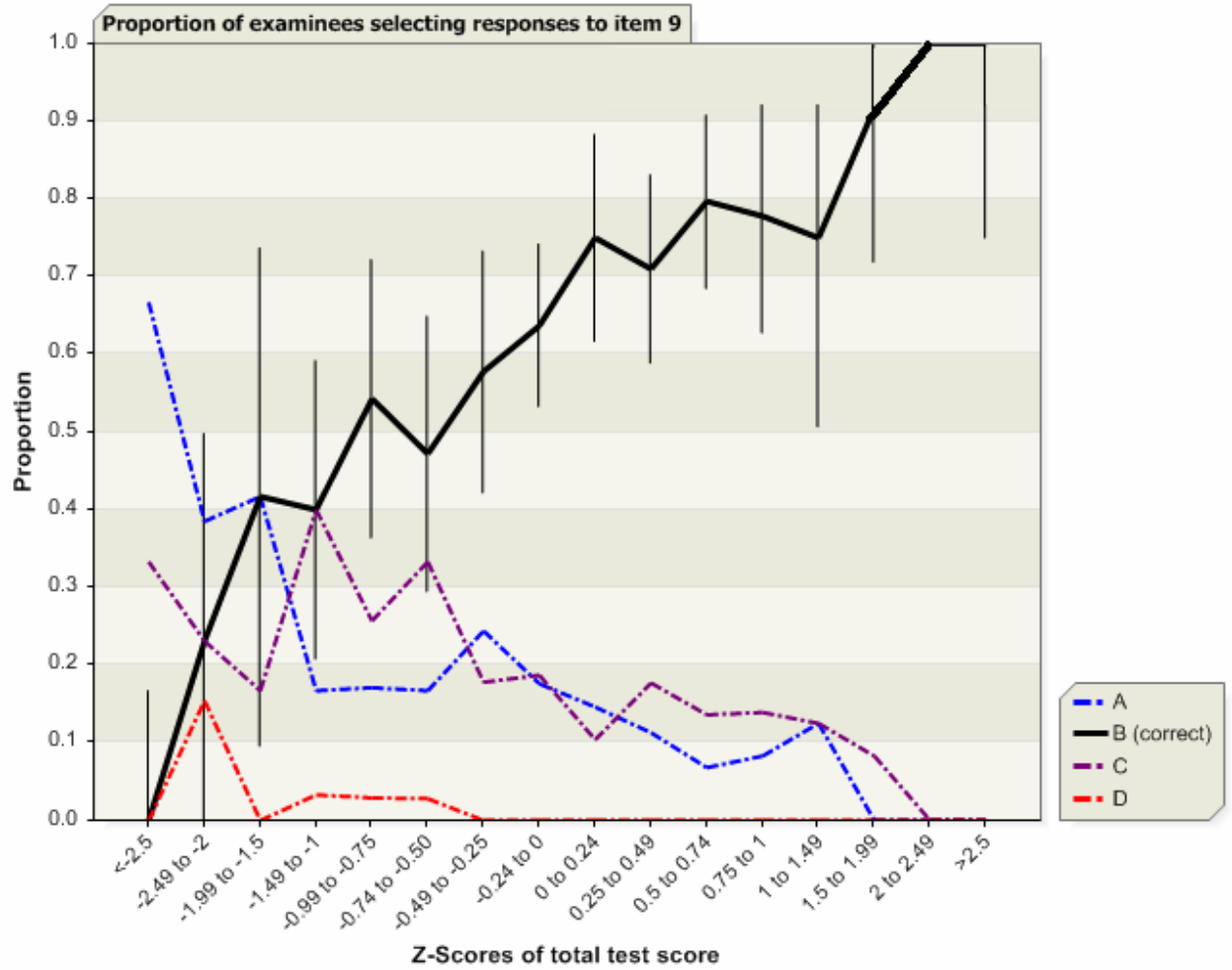
Group breakdowns

	Total	Top	Mid.	Low	TTS
No Answer	0.000	0.000	0.000	0.000	N/A
A	0.154	0.068	0.158	0.225	55.544
B	0.648	0.822	0.675	0.434	63.825
C	0.188	0.110	0.166	0.302	56.927
D	0.010	0.000	0.000	0.039	42.200

The information contained in the group breakdowns table is graphed in the item performance plot, the fourth section of the detailed item specific statistical report. Figure 4 displays the item performance plot for the item we just examined through the group breakdowns table. The Y-axis of the plot, labeled "Proportion," shows the proportion of students selecting each alternative. The X-axis, labeled "Z-scores of total test score," displays the normalized representation of the students' total test scores. In other words, the lowest scores on the test are plotted on left side of the X-axis, the middle range scores on the middle area, and the highest scores on the right side. We can see how the proportion of students who select the correct response (B) increases as we move from the left side of the graph to the right side of the graph (as the total test scores increase, so do the proportion of students selecting the correct answer). The opposite pattern is seen for the incorrect alternatives: as we move from the left of the graph to the right of the graph the proportion of students selecting each alternative decreases (as the total test scores increase, the proportion of students selecting the incorrect alternatives decrease). This graph allows us to visualize the proportion of students of different performance levels responding to each alternative. The steeper the line for the correct alternative, the more discriminating the item and, therefore, the higher the discrimination statistics. This is because a discriminating item makes sharp distinctions between students at different performance levels.

The black vertical lines surrounding the correct response option represents 95% confidence bars. The bars represent the range in which the "true" proportion of students selecting the correct response would fall 95 times out of 100. Factors such as the number of students will affect the width of the confidence bars (i.e., fewer students results in wider bars). This is because with fewer students, one is less confident that the "true" plot of the line will fall within a specific range, so the range is wider.

Figure 4. An example of an item performance plot produced by Integrity.



We will now look at some practical examples of methods to evaluate the performance of assessment items using Integrity.

A possible mis-keyed item

Figure 5 displays a possible mis-keyed item. We can see from the group breakdown table that only 7.6% of students selected the response option A, which was identified in the key file as the correct response option. Further, we can see that very few top-performing students selected response option A (3.4%). On the other hand, response option C drew 38.9% of the entire group of students, with 50.8% of the top-performing students selecting it. If we look at the item performance plot, we can quickly see this pattern, as the line for response option A trails at the bottom of the plot (i.e., almost no students selected this response option), and response option C looks like a possibility for the correct answer. In this situation, appropriate steps would be to examine closely the content of the item to determine which alternative is the correct alternative. If the key file was incorrect (e.g., due to a typo), fix the key and resubmit the job to Integrity. This will correct the discrimination statistics, increase the test mean, and should improve the

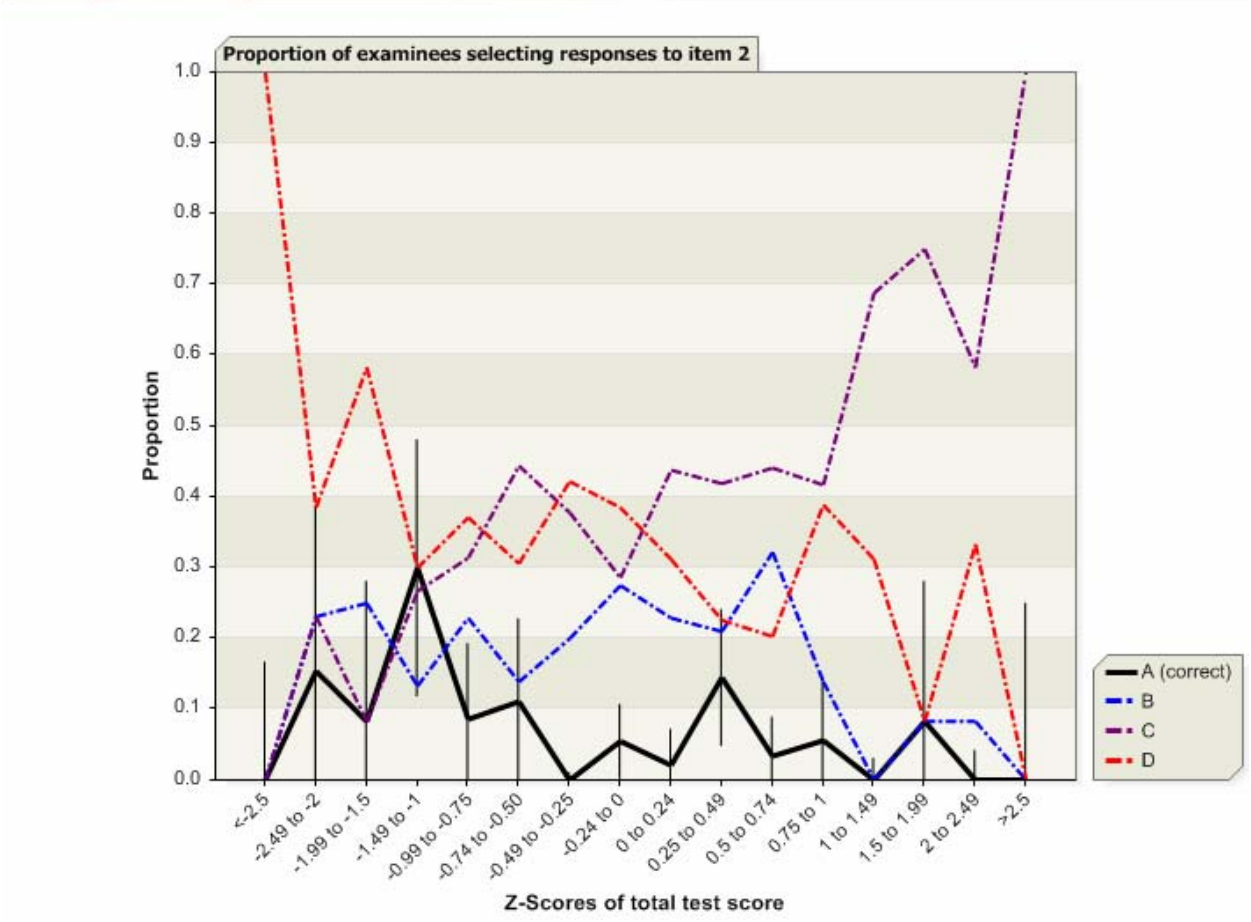
test overall (e.g., the reliability statistics may increase). If your analysis of the question shows that the response option A should be the keyed correct alternative, consider deleting this item from the test as it does not perform well for this group of students. This could occur for a variety of reasons: perhaps the material was too difficult for these students, perhaps students did not learn the concept being tested thoroughly enough, perhaps there is some ambiguity in the way the question is written, etc.

Figure 5. An example of a possible mis-keyed item identified by Integrity.

Group breakdowns

	Total	Top	Mid.	Low	TTS
No Answer	0.000	0.000	0.000	0.000	N/A
A	0.076	0.034	0.060	0.147	55.974
B	0.209	0.178	0.238	0.178	60.467
C	0.389	0.508	0.377	0.302	64.085
D	0.326	0.280	0.325	0.372	58.970

Item performance plot



An example of an item with high discrimination

Figure 6 displays an example of an item with high discrimination. The discrimination statistics (CPBR) for this item is 0.428, above the threshold of 0.350 that Integrity uses to indicate highly discriminating items. The summary statements reflect that this item is performing well statistically within a specific difficulty range. By examining the group breakdowns table, we see that the proportion of students selecting the correct response option is much greater for the high group than for the middle group, and much greater for the middle group than for the low group. If we examine the TTS column, we see that the total test scores for students who selected the incorrect response options “A,” “C,” and “D” are much less than the total test score for students who selected the correct response option of “B.” The item performance plot visually represents this information in that the slope or steepness of the correct response line is pronounced.

The third summary statement does flag a potential problem with the item in that alternative “D” draws a very small proportion of students. A review of the content for this alternative may be useful.

Figure 6. An example of an item with high discrimination identified by Integrity.

Item 26 - Key = B , Difficulty = 0.771 , Discrimination (CPBR) = 0.428

→ This item is of lesser difficulty.

→ This item performs very well statistically.

→ 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.

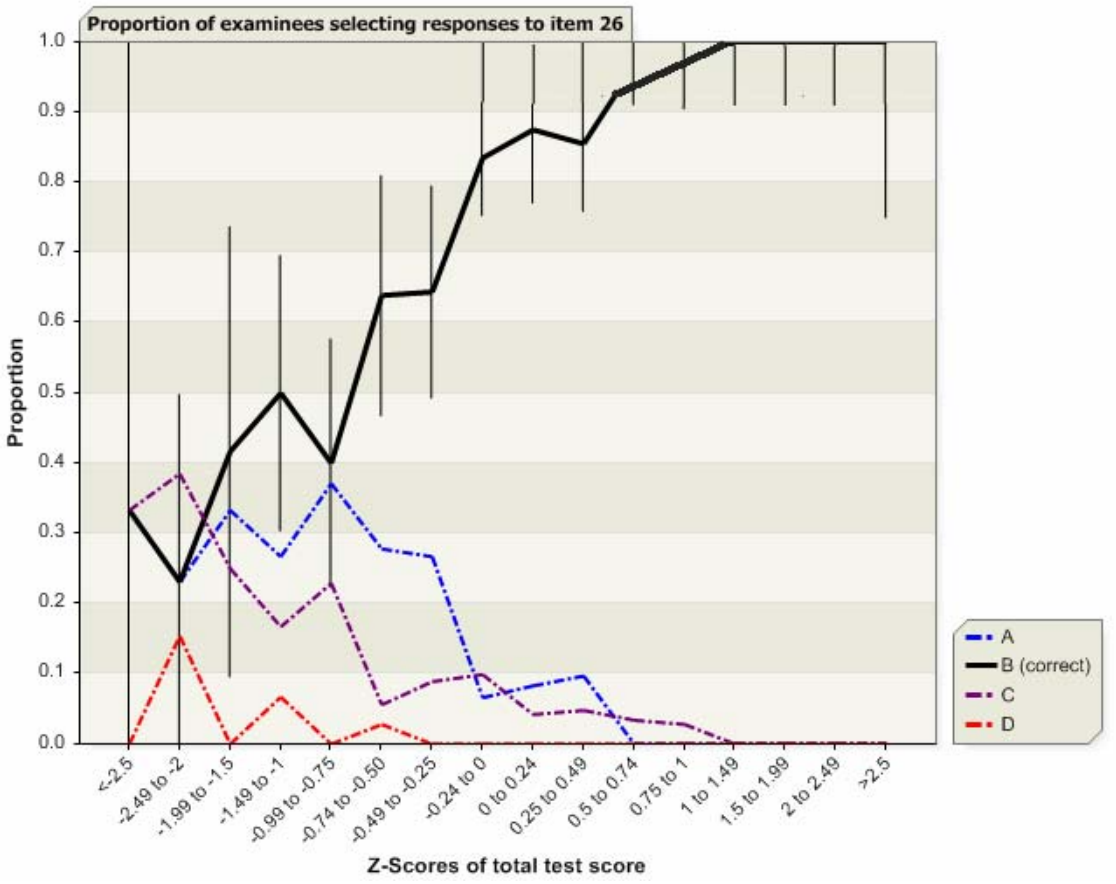
Correlation coefficients

Biserial correlation coefficient = 0.496	**Point-biserial correlation coefficient** = 0.457
Corrected biserial correlation coefficient = 0.464	**Corrected point-biserial correlation coefficient** = 0.428

Group breakdowns

	Total	Top	Mid.	Low	TTS
No Answer	0.000	0.000	0.000	0.000	N/A
A	0.131	0.000	0.106	0.302	51.493
B	0.771	0.992	0.819	0.473	63.970
C	0.088	0.008	0.075	0.186	51.733
D	0.010	0.000	0.000	0.039	41.600

Item performance plot



An example of an item with low discrimination

Figure 7 displays an example of an item with low discrimination. The discrimination statistic (CPBR) for this item is 0.095, far below the threshold of 0.225 that Integrity uses as indicating low discrimination. The second summary statement for the item indicates that the item is displaying low discrimination and states that “Examinees of low ability should have a much lower probability of answering an item correctly than do examinees of high ability.” If we look at the group breakdowns table, it shows that for the correct response option (C), middle and low students have similar likelihoods of selecting the correct response. We would expect that fewer lower-performing students would select the correct response than middle-performing students. The item performance plot shows this pattern graphically: the steepness, or slope, of the correct response option is a visual measure of discrimination. The steeper the correct response option line is, the more discriminating the item. This is because we expect that a lower proportion of low-performing students would select the correct response that the middle-performing students, and that a lower proportion of middle-performing students would select the correct response than the top-performing students. We see that for this item, the line is not steeply rising as we move from the left side of the X-axis to the right side of the X-axis (i.e., as we move from low-performing to high-performing students). It may be useful to examine the content of this item to determine whether the wording was ambiguous, or if something in the content was inadvertently giving away the answer to the low-performing students.

We see that for this item, approximately 35-45% of students at a very low performance level (to the far left on the X-axis) are selecting the correct alternative to the item. This is unusual — we would typically expect very few students at such a low performance or ability level to select the correct answer.

The TTS column of the group breakdowns table shows that the total test scores for student who selected the incorrect response options B (60.467) and D (58.031) are close to the total test scores for students who selected the correct response option, C (62.372). This is not expected because students who choose the incorrect answers to an item should have significantly lower overall total test scores than students who choose the correct answer to an item. The third summary statement points out an additional potential problem with the item, that the incorrect response option A draws a very small proportion of students. This is visually illustrated in the item performance plot. It may be useful to revise this alternative to attempt to draw more students.

Figure 7. An example of an item with low discrimination identified using Integrity.

Item 86 - Key = C , Difficulty = 0.594 , Discrimination (CPBR) = 0.095

→ 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 A for this item. Consider reviewing the content of this alternative to determine if it can be revised to attract more examinees.

Correlation coefficients

Biserial correlation coefficient = 0.169	**Point-biserial correlation coefficient** = 0.136
Corrected biserial correlation coefficient = 0.118	**Corrected point-biserial correlation coefficient** = 0.095

Group breakdowns

	Total	Top	Mid.	Low	TTS
No Answer	0.000	0.000	0.000	0.000	N/A
A	0.014	0.000	0.019	0.016	53.286
B	0.205	0.178	0.223	0.194	60.467
C	0.594	0.695	0.570	0.550	62.372
D	0.188	0.127	0.189	0.240	58.031

Item performance plot