Quick start guide to interpreting collusion detection results from Integrity

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

Rational for conducting collusion detection analyses

The rational for the use of a system to identify students who may have engaged in collusion (i.e., cheating – answer copying) when taking tests lies fundamentally in the idea that creating a "level playing field" makes testing more fair for all students. If some students are cheating on tests and obtaining higher grades than they should, then by comparison, students who do not cheat may be disadvantaged. The pressure for those honest students to cheat on tests then increases. Cheating has major implications for students, the public, businesses, and education centers in that they may be hiring/selecting/passing students who are not properly qualified.

From a test validity perspective, cheating on tests has impacts on the inferences that can be made from the results. Test validity is generally defined as how well the test measures what it is supposed to measure. If students are cheating on tests, the validity of the test results is eroded because a "true" reflection of the students' ability (or aptitude, or achievement, etc.) is not being measured. As a result, the inferences one can make from the test results will be limited because one is no longer as certain as to what the scores mean. Academic centers that do not monitor and address issues of cheating may receive negative reputations from employers who eventually hire their graduates and find that the performance of those graduates is below acceptable standards.

In the popular book *Cheating on Tests, How to Do It, Detect It, and Prevent It,* Dr. Gregory Cizek (1999)[1] states that nearly all research conducted on the prevalence of cheating, at all levels of schooling but especially at the post-secondary level, indicates that "cheating is rampant" (pg. 13) and that the number of students who admit to cheating has increased in the last 30 years. As a result, it would appear that there is a need for statistical tools, such as those offered through Integrity, that evaluate the academic integrity of test takers and test administration centers.

Role of statistical collusion detection methods

Professional opinions regarding the use of statistical methods to detect cheaters vary. Some argue that statistical methods to detect cheaters should only be used if a pair of students is identified through some other means as being suspicious (e.g., a test proctor

[1] Cizek, G. J. (1999). *Cheating on Tests, How to Do It, Detect It, and Prevent It.* Lawrence Erlbaum Associates: Mahwah, New Jersey.

believes that he/she saw suspicious behavior occurring between two students during an examination). As all statistical and non-statistical methods have false positive rates (identifying a pair of cheaters who did not actually cheat) this school of thought advocates that using statistical methods alone to flag students may be inappropriate. Others argue that using statistical tests to identify cheaters in a more exploratory manner is appropriate but should not be the sole source of information. Statistical tests should be only one piece of evidence used when investigating a pair of students with suspicious answer responses.

Clearly a balanced approach is necessary when addressing such a complex and delicate area as test collusion. Simply broadcasting the systematic use of Integrity at an institution may reduce the number of examinees who consider engaging in collusion. In addition, it may be prudent to use Integrity as one of many tools (such as the systematic use of test proctors) in the identification and verification of possible cheaters. Most educational institutions have protocols in place to investigate allegations of cheating. As such, using information provided by Integrity as one piece of evidence in the institutions existing investigative process is the recommended approach. It is not recommended that the collusion detection results of Integrity be used as the only source of evidence in an investigation of cheating. For an extensive exploration of the issues surrounding cheating on tests, the popular book by Dr. Gregory Cizek (1999) is recommended.

The main purposes of the collusion detection features of Integrity are to: a) provide information on the integrity of writing centers (i.e., are the test-taking processes secure), b) provide information on the academic integrity of examinees via their responses to multiple-choice type questions.

Collusion reports produced by Integrity

The collusion detection analysis is an optional analysis within Integrity. When submitting test data to Integrity, you will be asked via the step-by-step wizard (see Figure 1) whether you would like to have the collusion detection analysis run. If you select the collusion detection analysis, you will be asked to choose one of three options for how the analysis should be conducted (see Figure 2):

1) All students – this option conducts collusion analysis on all students who took the test. It compares the responses on each question for all possible pairs of students
2) Specific students on all students – this option compares the responses of students that you specify (e.g., suspect) with the responses of all other students who took the test
3) Only investigate specific students – this option compares the responses of only the students that you specify (e.g., suspect) with each other

Figure 1. An example of selecting collusion detection analysis from the data submission wizard within Integrity.

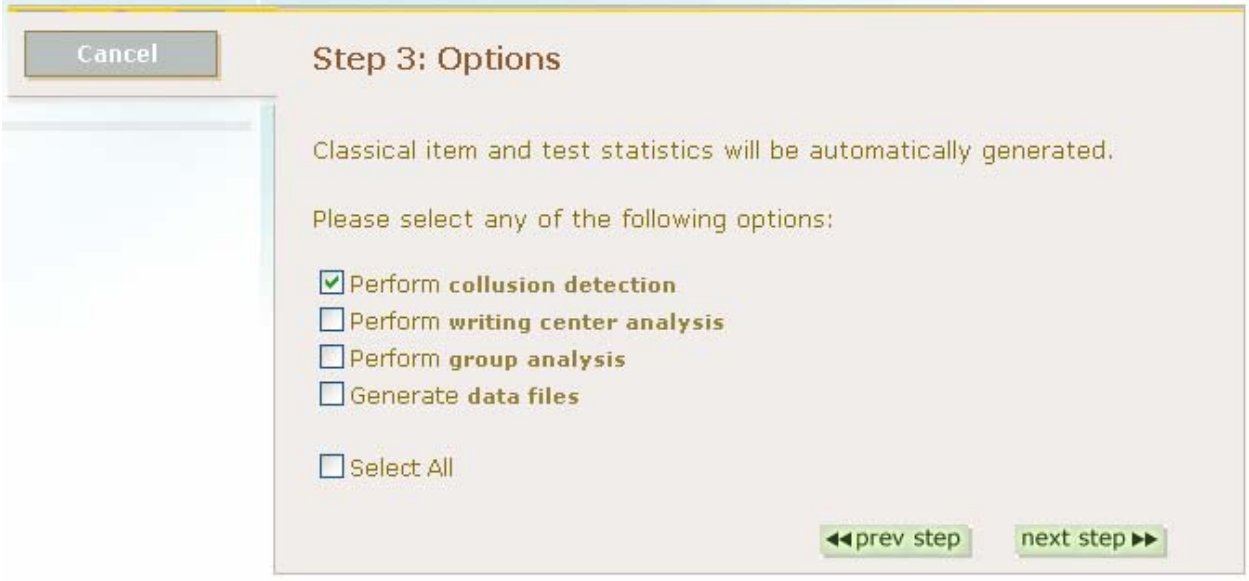
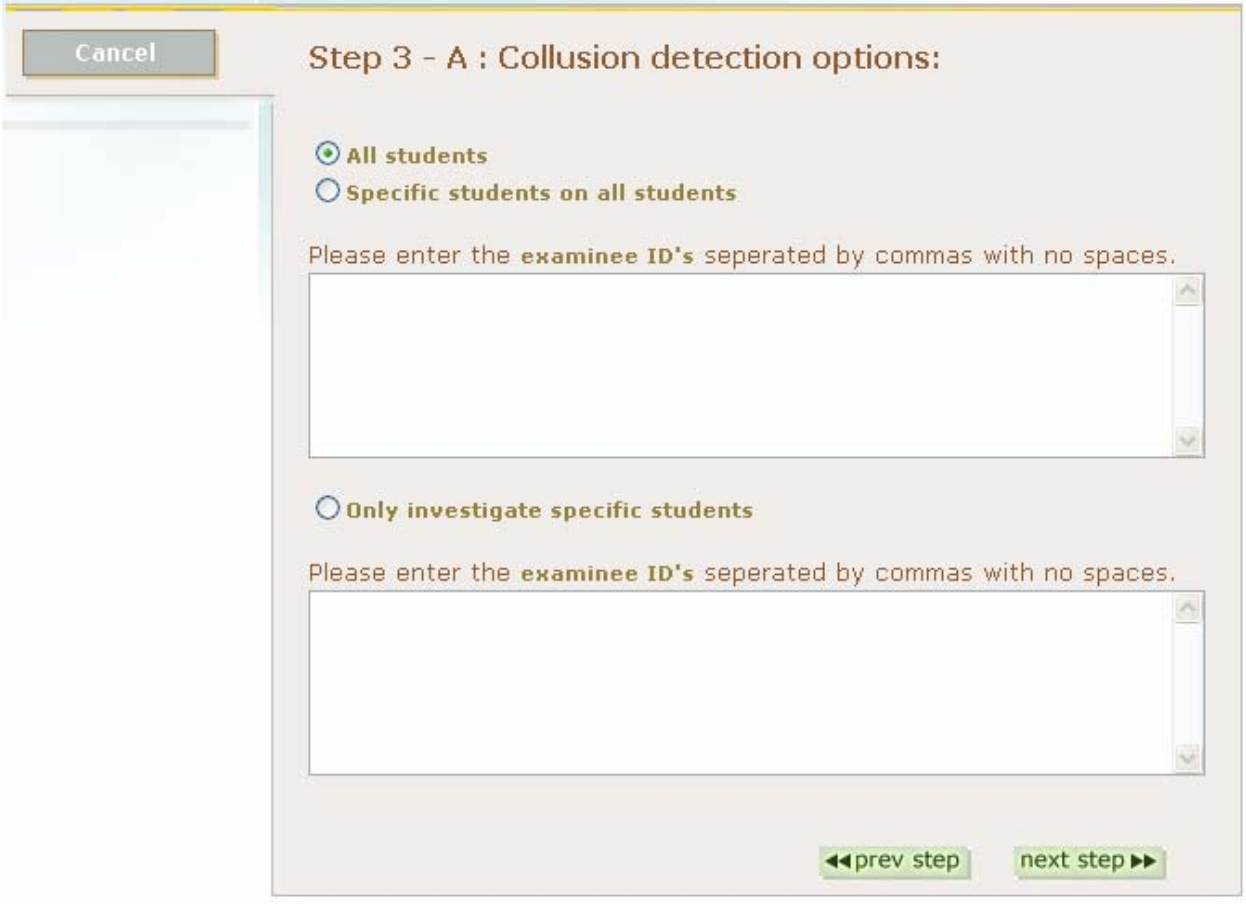


Figure 2. An example of the three collusion detection analysis options within Integrity.



After you have selected the appropriate collusion detection options and finished submitting a job (test) to Integrity, the collusion detection reports are generated. Integrity utilizes five statistical collusion detection methods that have been vetted in the academic research literature. Those interested in learning more about each method are encouraged to review the descriptions of each method within the application by clicking on the terms or visiting the glossary.

An example of this collusion detection summary report, which is presented in the Executive summary report, is shown in Figure 3. The example shows three pairs of examinees that have been identified via at least one of the five collusion detection methods. The writing center information (an optional piece of information that can be contained in the data file) is displayed in order to ensure that the pair of examinees took the test at the same writing location. For each of the five collusion detection methods, four categorizations of statistical certainty are possible. Statistical certainty indicates the level of statistical confidence that a pair of examinees have engaged in collusion. Typically, the more rare the pattern of responses for a pair of examinees, the greater the statistical certainty that the pair of examinees engaged in collusion. The four levels of statistical certainty used by Integrity are:

1) N/A = the pair of examinees did not meet the statistical threshold set by Integrity, for that particular method, to be identified as potentially engaging in collusion
2) Low = there is a “Low” statistical confidence, based on that particular method, that a pair of examinees engaged in collusion. One can expect more false positives at this level than at other levels
3) Moderate = there is a “Moderate” statistical confidence, based on that particular method, that a pair of examinees engaged in collusion. One can expect some false positives at this level, but fewer than at the “Low” level
4) High = there is a “High” statistical confidence, based on that particular method, that a pair of examinees engaged in collusion. One can expect very few false positives at this level

Figure 3. An example of a collusion detection summary report produced by Integrity.

Collusion detection report

→ 3 pairs of examinees have been identified by the collusion detection analysis.

	Examinee ID	Writing center	B-Index	PAIR1	PAIR2	MESA	g2
Examinee pair 1	10405047 10405048	1111 1111	High	High	N/A	Low	High
Examinee pair 2	10505049 10505050	1111 1111	High	N/A	N/A	Moderate	High
Examinee pair 3	105050149 105150150	3333 3333	Low	N/A	N/A	N/A	Low

It would be considered more probable for pairs of examinees who have “High” statistical certainty for collusion detection methods to have engaged in collusion than for pairs of examinees who have “Low” statistical certainty for collusion detection methods. The

thresholds for these statistical confidence levels were set based on theoretical attributes of each method, peer-reviewed research literature, and estimated false-positive versus true-positive rates. The thresholds were set to maximize true-positive identification (examinees who engaged in collusion and are identified) and minimize false-positives (examinees who did not engage in collusion and are identified). The threshold for a "Low" statistical certainty has been set at a very conservative level and therefore very few false positive cases can be expected at this statistical certainty level. As such it would be prudent to not discount examinees identified at this statistical certainty level.

In the example shown in Figure 3, "Examinee pair 1" is identified with a high statistical certainty using the B-index method, the PAIR1 method, and the g_2 method. The MESA method identified the pair of students at a "Low" statistical certainty and the PAIR2 method did not identify the examinee pair at all. In contrast, "Examinee pair 3" was identified only by the B-index and g_2 methods at a "Low" statistical certainty. In this example, it may be prudent to interpret the collusion detection results for "Examinee pair 3" with caution as the overall statistical certainty for that pair of students is low.

Because each of the five collusion detection methods uses a different approach to identify collusion, if even one approach identifies a pair of students as potentially engaging in collusion, even at a "Low" statistical certainty, that examinee pair should be scrutinized.

The collusion detection report main page shows a summary of the pairs of examinees identified using each collusion detection method. By clicking into the "Collusion detection" report link, you can access this main page. The main page presents a table that presents the statistical certainty and statistical values for each collusion detection method for each pair of examinees identified (see Figure 4).

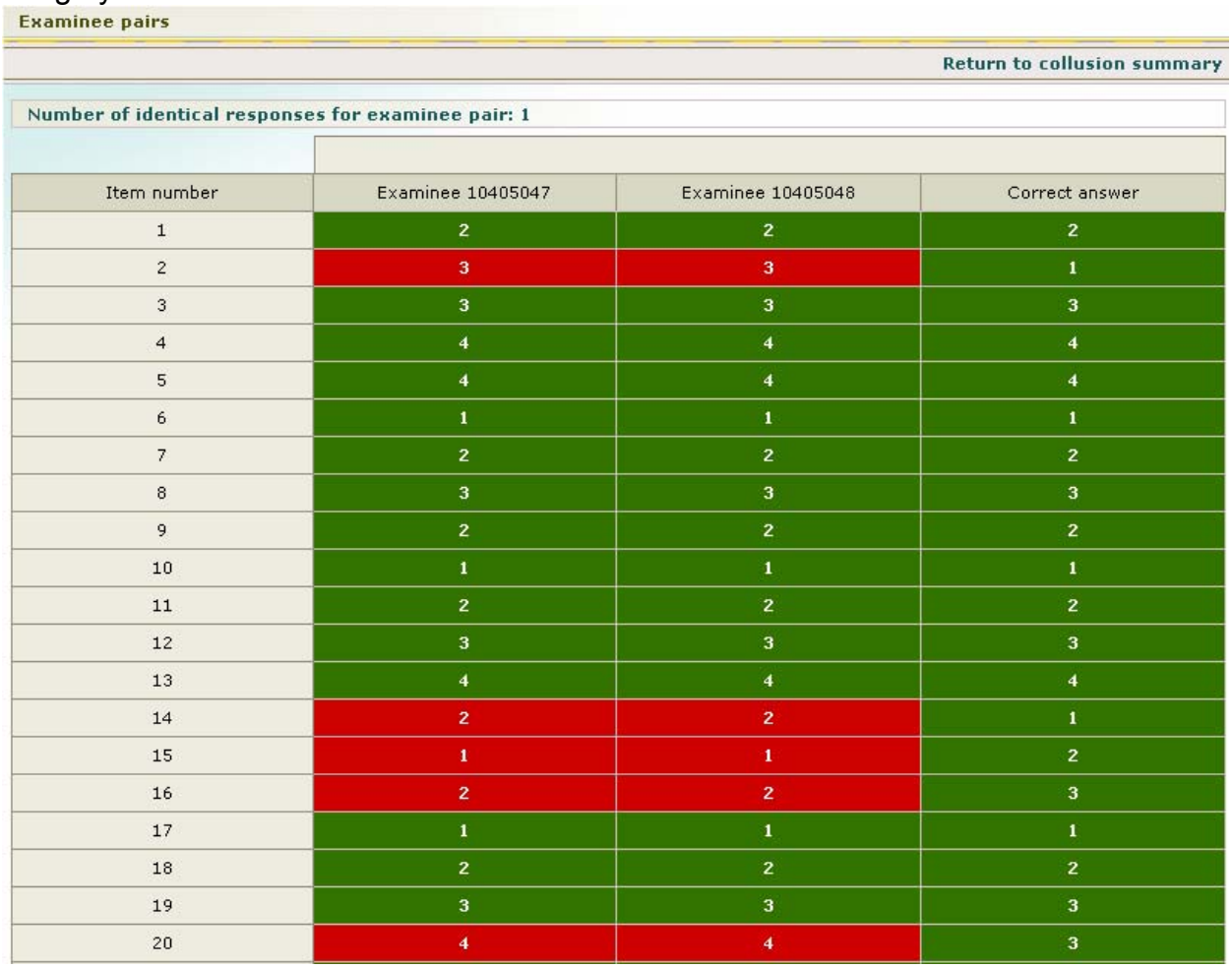
Figure 4. An example of a collusion detection main report table produced by Integrity.

Collusion detection report

	Examinee ID	Writing center	B-Index	PAIR1	PAIR2	MESA	g2
Examinee pair 1	10405047 10405048	1111 1111	High 11.707	High 2900	N/A	Low 1.568E-008	High 8.678 8.678
Examinee pair 2	10505049 10505050	1111 1111	High 10.231	N/A	N/A	Moderate 6.979E-011	High 9.181 9.094
Examinee pair 3	105050149 105150150	3333 3333	Low 7.738	N/A	N/A	N/A	Low 6.063 6.089

By clicking on the examinee pair (e.g., "Examinee pair 1") you can access a table that lists the item responses for each examinee. Green cells represent correct responses to items and red cells represent incorrect responses to items. The resemblance of the responses to items between the pair of examinees allows for an intuitive review of how similar, and therefore how rare, the pattern of responses for that pair of examinees is. An example of the examinee pair response table is presented in Figure 5.

Figure 5. An example of a section of an examinee pair response table produced by Integrity.



Examinee pairs

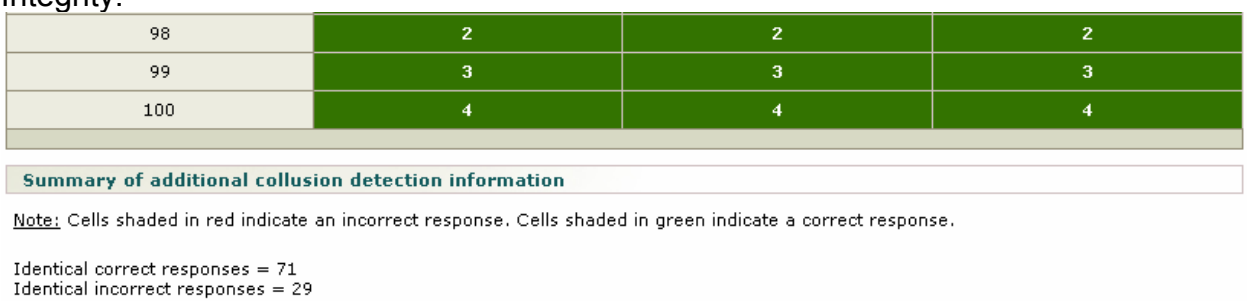
Return to collusion summary

Number of identical responses for examinee pair: 1

Item number	Examinee 10405047	Examinee 10405048	Correct answer
1	2	2	2
2	3	3	1
3	3	3	3
4	4	4	4
5	4	4	4
6	1	1	1
7	2	2	2
8	3	3	3
9	2	2	2
10	1	1	1
11	2	2	2
12	3	3	3
13	4	4	4
14	2	2	1
15	1	1	2
16	2	2	3
17	1	1	1
18	2	2	2
19	3	3	3
20	4	4	3

At the bottom of the examinee pair response table is a summary of the number of identical correct and incorrect responses for the pair of examinees, an example of which is presented in Figure 6. In this example, we see that the examinee pair selected 71 of 100 identical correct responses and 29 of 100 identical incorrect responses. In other words, all of their responses to all of the items composing the test were identical.

Figure 6. An example of the examinee pair response summary information produced by Integrity.



98	2	2	2
99	3	3	3
100	4	4	4

Summary of additional collusion detection information

Note: Cells shaded in red indicate an incorrect response. Cells shaded in green indicate a correct response.

Identical correct responses = 71
Identical incorrect responses = 29

By using the dropdown menu above the collusion report, you can navigate into the detailed collusion reports for each of the five methods. An example of the Angoff's B-index results report is presented in Figure 7. The report first lists the B-index results for each examinee pair identified using the B-index method, followed by a histogram of the results. We can see that the B-index value for "Examinee pair 1" is 11.707. By looking down to the histogram, we can see where this "event" occurs in relation to all other B-index values for all other pairs of examinees analyzed. The arrows on the histogram show the location of the 11.707 value, far to the right of the majority of other values. In fact, for the Angoff method, the mean of all values is approximately zero with a standard devaition of one (as displayed in the legend window), indicating that a value of 11.707 (almost 12 standard deviations from the mean) is an extremely rare event resulting in a "High" statistical certainty. By way of comparison, "Examinee pair 3" is shown to have a B-index value of 7.738, far to the right of the majority of B-index values but not as far as the value of 11.707 for "Examinee pair 1." As such, this examinee pair is labeled as a "Low" statistical certainty. Please note that the "n" in the legend box refers to the number of pair comparisons made during the collusion analysis, not the number of students who took the test.

Figure 7. An example of the Angoff's B-index results report produced by Integrity.

Detailed collusion detection report (all examinees) for job: Psychology 101 session A1

Angoff's B-index results

Angoff's B-index collusion detection method

Examinee pairs that exceeded collusion detection threshold for this method:

Examinee pair 1

Examinee ID	Writing center	B-index	Statistical certainty
10405047 10405048	1111 1111	11.707	High

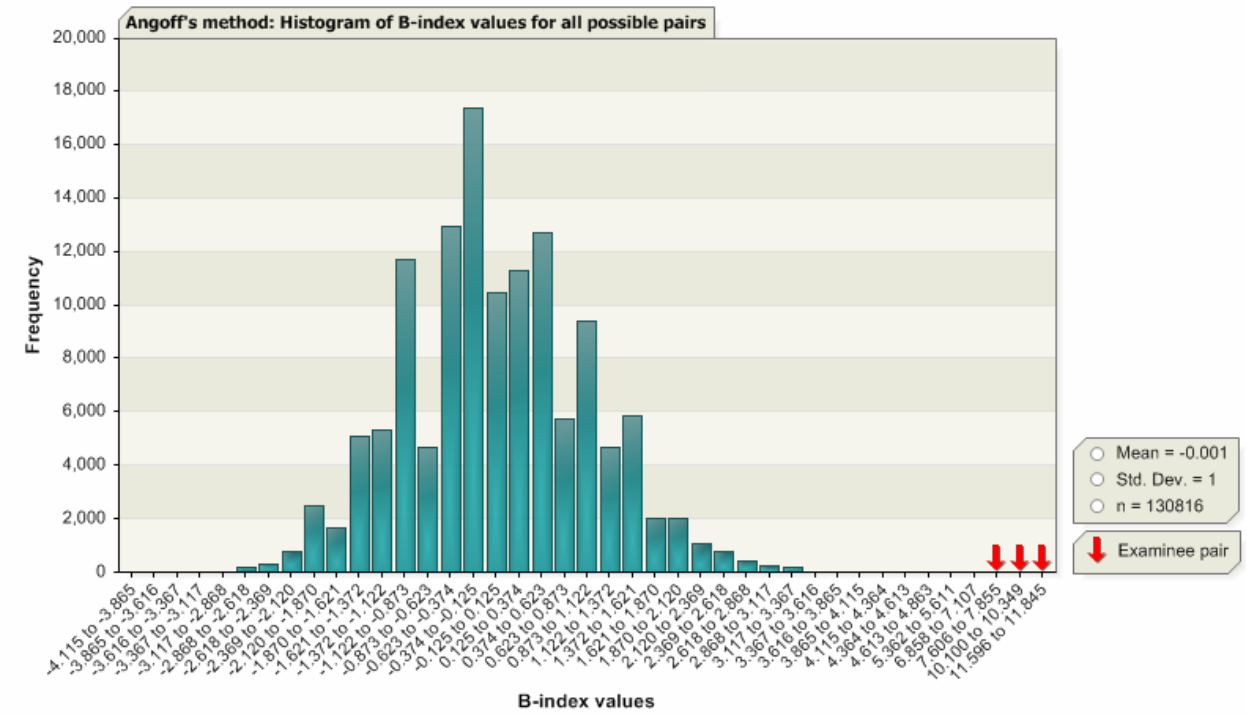
Examinee pair 2

Examinee ID	Writing center	B-index	Statistical certainty
10505049 10505050	1111 1111	10.231	High

Examinee pair 3

Examinee ID	Writing center	B-index	Statistical certainty
105050149 105150150	3333 3333	7.738	Low

Histogram of B-index values for all possible pairs



In order to obtain more specific information regarding how rare a collusion detection value is, one can look on a standard normal distribution table to find approximate probability values. For example, in the example above "Examinee pair 3" has an Angoff

B-index value of 7.738. The graph in Figure 4 shows that the distribution of all the B-index values for all pairs of examinees is quite normally distributed and has a mean of approximately zero and a standard deviation of one. As such, a B-index value of 7.738 indicates that the rarity of the answer similarity analysis for "Examinee pair 3" is almost 8 standard deviations from the mean. Table 1 presents the approximate probability values of events far to the right of a standard normal distribution. The probability for a standard deviation of "1" is 0.1587, or approximately one in six (i.e., 15.87/100), the probability for a standard deviation of "2" is 0.0228, or approximately one in 50 (i.e., 2.28/100). As the number of standard deviations to the right of the distribution increases the probability become less and less. An event at 8 standard deviations from the mean is extremely rare, in the order of billions to one. Of course, an event almost 12 standard deviations from the mean (Examinee pair 1) is even more rare.

Table 1. Approximate probability values to the far right of the standard normal distribution.

Number of standard deviations to the right of the mean	Approximate probability of event
1	0.158700
2	0.022800
3	0.001350
4	0.00003167*
5	2.867 E-7
6	9.866 E-10
7	1.280 E-12
8	6.221 E-16
9	1.129 E-19
≥9.5	≤1.049 E-21

*In scientific notation this value = 3.167 E-5

The results of other collusion detection methods can be interpreted in a similar manner to the example we just reviewed. The MESA method values are unique, however, in that the MESA values are probability values. As such, if a pair of students is identified as having an MESA value of "1.155E-8," this value can be interpreted directly as the approximate probability of the event occurring.

In a formal investigation process at an educational institution, it may be prudent to call upon the services of an expert in the area of statistics in order to provide a very comprehensive explanation of probability estimation using collusion detection statistics.

