CONFIDENTIAL

Agreement Analysis Report

When an organization requests an Agreement Analysis, the provides baseline information, in addition to statistical results for the identified pair(s) of examinees. Baseline information is derived from the responses of pairs of examinees for whom there is no report of copying or collaboration. This information is then used to establish empirical baseline distributions of the z-statistic representing the level of expected agreement on items answered incorrectly by pairs of examinees. Prior to the calculation of the z-statistic for any pair of examinees, a comparison group of examinees deemed appropriate for a particular analysis and for whom there is no reason to question the validity of their scores is identified. The comparison group usually shares characteristics with the pair of examinees in question. The item responses of examinees in the comparison group are used to determine the empirical estimates of the expected probability of wrong answers for items on the test. In general, larger comparison groups deemed appropriate for a particular analysis provide more stable test statistics.

Nineteen (19) examinees <u>other</u> than the two examinees identified above who took the same form of the were identified as appropriate for inclusion in the Comparison Group. Since this number of comparison group examinees is relatively small (n = 19) for statistical purposes, the use of these examinees alone as a comparison group may have compromised the stability of the test statistics. Therefore, to increase the sample size for statistical purposes, in this application of the Agreement Analysis a larger comparison group of examinees was identified. Examinees who took the same form of the Examination as the examinee pair in question and who tested were identified as appropriate for inclusion in this group. This comparison group included sixty-seven (67) examinees, forty-eight (48) examinees from other medical schools and nineteen (19) examinees from your school.

Information on the two examinees in question, the comparison group, empirical baseline distribution and the results of the analyses are summarized on the following pages.

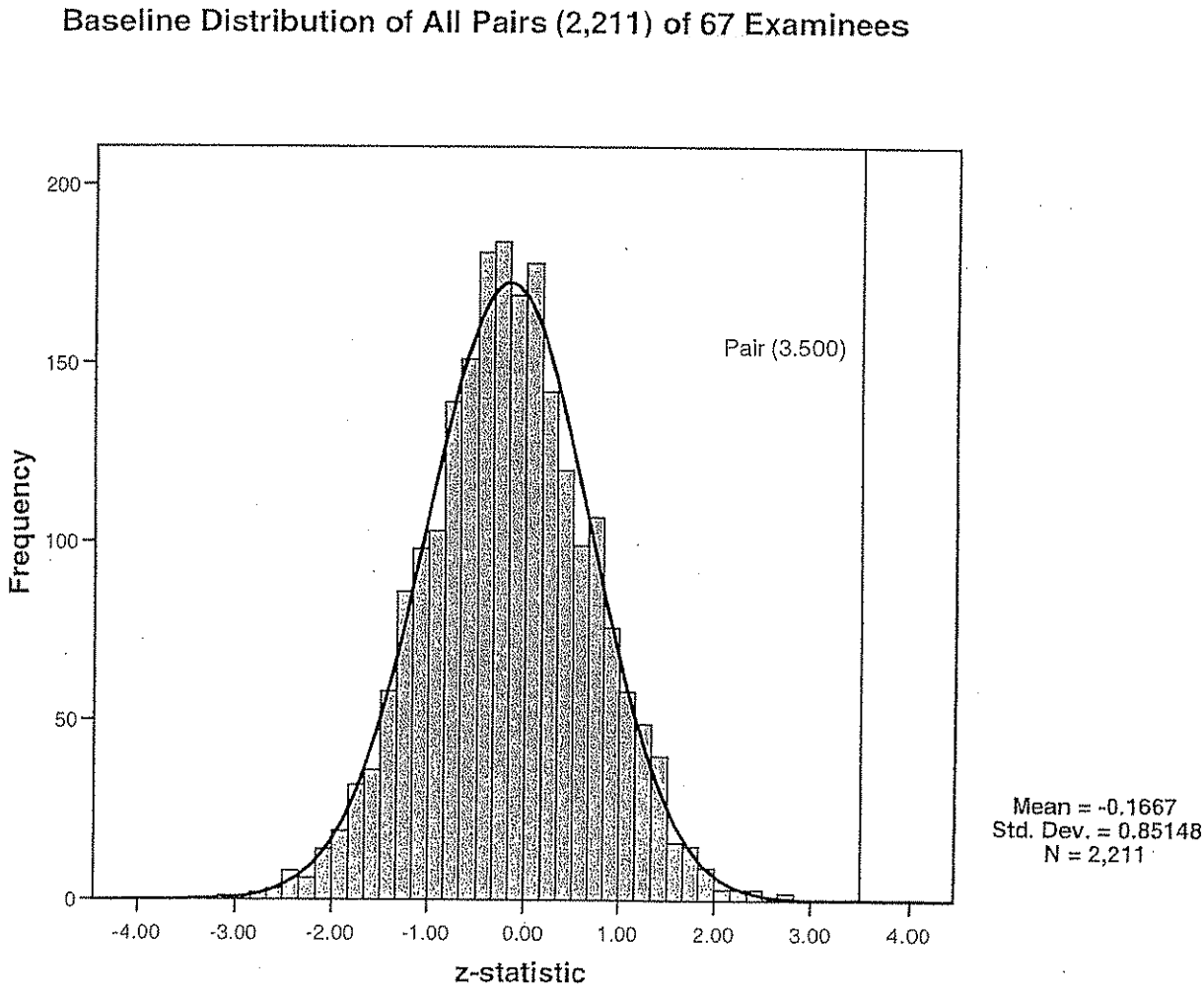
Description of Agreement Analysis

Comparison Group

The Comparison Group was comprised of 67 examinees (including 19 examinees from your school) who took the ! Examination

Empirical Baseline Distribution

The baseline group was all possible pairs (2,211) of the sixty-seven (67) examinees in the Comparison Group. The empirical distribution of the test statistic (z) for all 2,211 possible pairs of examinees in the comparison group is shown below, along with a superimposed normal curve. For the empirical distribution below, the minimum z observed was –3.05 and the maximum z was 2.79.



Using a standard table of areas under the normal distribution curve, the probability associated with a z greater than 3.09 is less than 1 in 1,000 and the probability for a z greater than 3.72 is less than 1 in 10,000

Agreement Analysis results for [redacted] Candidate X

Candidate X had 72 correct responses and [redacted] had 79 correct responses out of a total of 100 items on the Examination. Candidate X (suspected copier) was paired with [redacted] (source) and their incorrect responses were analyzed using the comparison group described above.

Number of Incorrect Answers		Count of Items Answered Incorrectly by Both Examinees	Count of Joint (Same) Incorrect Answers by Both Examinees	Percentage of Joint (Same) Incorrect Answers		Calculated Z*
				Observed Agreement	Expected Agreement	
28	21	14	13	93%	46%	3.50

* Agreement Analysis Z-statistic if candidate X is assumed to be the copier.

A z value of 3.50 was calculated for the difference between the observed percentage of agreement (93%) and the expected percentage of agreement (46%). None of the 2,211 z values from the empirical baseline distribution using all possible pairs of examinees in the Comparison Group was greater than 3.50. This would imply that, assuming the sixty-seven (67) comparison group examinees who took the Examination responded independently, the probability of obtaining a z greater than 3.50 by chance for the empirical baseline distribution is less than 1 out of 10,000 ($p < .0001$). It should also be noted that, if the assumptions on which this analysis is based are met, the probability associated with a z greater than 3.50 using a standard table of areas under the normal distribution curve is less than 3 out of 10,000 ($p < .0003$), thus further substantiating the results above.

Agreement Analysis

The Agreement Analysis compares the degree of agreement that is observed between the wrong answers of two examinees with the degree of agreement that would be expected to occur among two randomly chosen examinees taking the same test independently. The Analysis uses only those test items that both examinees in the pair answered incorrectly. It can be useful in instances of irregular behavior involving observed incidents when one examinee may be copying the responses of another examinee, two or more examinees are sharing responses, or there has been opportunity for collaborative development of an answer key to an examination obtained prior to administration.

Rationale

A statistical model that may be employed in this analysis is the normal approximation of the binomial expansion. This model is generally accepted by statisticians as a valid means for calculating the probability of obtaining a specific number of events from a series of independent trials when the probability that the event will occur by chance is constant and can be specified for each trial. It has been noted that, in practice, the probability that two examinees working independently, will give identically incorrect answers to each item both answered incorrectly, is rarely constant. The violation of this assumption makes the statistical test less sensitive, in favor of the examinees in question. The assumption of independence of items is also subject to debate in any practical application of this model and serious violation of this assumption could operate against the pair of examinees in question.

When there are large numbers of examinees who are similar to the suspected pair, an empirically-based distribution of the statistic is often computed to estimate the probability of a chance event. In many instances, the empirical distribution closely approximates the normal distribution.

Procedure

To perform an Agreement Analysis, one first identifies a comparison group of examinees deemed appropriate for a particular analysis, about whom there is no reason to question the validity of their scores. The comparison group usually shares characteristics with the pair of examinees in question. A comparison group, for example, may be examinees enrolled in the same medical school who have taken the same form of the test for a similar purpose, those who have completed a comparable level of education and are taking the examination for a similar purpose, those who have attained a similar degree of medical training, or those who have achieved a similar score on the same form of the test as the pair in question. At times, the comparison group may include all examinees tested.

The responses of the comparison group provide the empirically determined, expected percentage of incorrect responses against which the incorrect responses of the pair is compared. In general, larger comparison groups deemed appropriate for a particular analysis provide more stable test statistics than smaller comparison groups.

The first step in the calculation of a test statistic (z) is to determine the number of items both examinees in a pair answered incorrectly (joint wrongs) and then, determine how frequently both examinees gave the identical wrong answers to these items. The expected number of wrong answers is based on the responses of the comparison group. For each of the joint wrong answers, the number of examinees in the comparison group who also answered the item incorrectly is determined and the number of examinees with the same wrong answer as that given by the examinee who is designated the source of the responses is calculated. The percentage derived from these two numbers is the expected (chance) probability that another examinee working independently would select the same wrong answer as that given by the source of the responses. The percentages are averaged for all joint wrongs.

A test statistic (z) is calculated to estimate the probability that the number of identical wrong answers given by the two examinees would have occurred by chance, when the two examinees answered these same items independently of each other. When the copier is identified, the z-statistic is based on the incorrect responses of the other examinee as the source of the responses. When collusion between examinees is suspected, a z-statistic can be calculated for each examinee in a pair, designating each examinee as the copier of responses and the other as the source of the responses. The general formula is:

$$z = \frac{(A - E)}{SD}$$

where:

Z	=	normal deviate
N	=	number of items answered incorrectly by both examinees (joint wrongs)
P	=	empirically estimated probability that two examinees working independently of each other would give identical incorrect answers to any one of the N items by chance alone
Q	=	1-P (empirically estimated probability that two examinees working together independently of each other would not give identical incorrect answers to any one of the N items by chance alone)
A	=	number of identical wrong responses among the joint wrongs
E	=	number of identical wrong responses expected by chance (PN)
SD	=	square root of NPQ.

The probability associated with the resulting z-statistic is obtained from a statistical table of areas under the normal distribution curve or from an empirical distribution of z-statistics from a baseline group. When the probability associated with a z-statistic is very small, one may reject the hypothesis that the agreement observed between the examinees occurred by chance alone.

Use and Interpretation of Results

When carries out an Agreement Analysis, it is with the understanding that the use of the analyses is the responsibility of the requesting organization or medical school. Caution is always advised when interpreting the results of these analyses.

In order to provide adequate protection against rejecting the null hypothesis (i.e., that the observed agreement between the examinees in question occurred by chance alone) when it is true, alpha levels that are more conservative than those commonly used in scientific investigations are used when interpreting the results of an Agreement Analysis. Alpha levels sometimes as small or less than 1 in 10,000 (.0001) are used to conclude that a statistically significant degree of agreement was found in any one analysis. However, a different alpha level may be selected based on other relevant information such as the availability of other types of evidence that an irregularity may have occurred, the potential impact on the examinee, etc.

If an investigation is initiated as a result of a report which identifies a copier-source pair of examinees, only one statistical test is required. However, if a more widespread degree of collusion is suspected, it may be necessary to conduct analyses for many pairs of examinees. In such cases, the alpha level may be adjusted for the fact that multiple statistical tests have been performed.

The provides baseline information in addition to the statistical results for the examinees suspected of irregular behaviors when an Agreement Analysis is requested by an organization or a medical school. Baseline information is derived from the responses of examinees, unlikely to have copied or collaborated with one other. Generally, this information is based on the degree of agreement (z-statistic) between the incorrect responses of all possible pairs of examinees included in the comparison group of the analysis performed on a suspected pair or pairs of examinees. The distribution of the test statistics, based on the incorrect responses of each examinee of a pair as the source of the responses separately, is used to assess the assumptions on which this technique is based and applicability of the technique to the test data.

Agreement Analysis is a statistical tool that can provide helpful, supporting information for the investigation of observed behaviors that may compromise the validity of examinee test scores. When an agreement is statistically significant, these analyses do not indicate what factor or factors led to an agreement which was observed, and there is always some probability (however small) that the observed agreement did in fact occur by chance. Therefore, the results of such analyses are not sufficient by themselves for arriving at a conclusion that an examinee or examinees engaged in irregular behavior. This conclusion is a judgment that may be supported but is never proven by the statistical analysis and one that should be made on the basis of all relevant information that is available.