Considerations when Interpreting Critical Thinking Skills Test Scores

The interpretation of the importance of group descriptive statistics is the responsibility of the client. For example, it is the responsibility of the client to decide how much of a gain from pretest to posttest in group mean scores or percentiles is large enough to be regarded as satisfactory or important for the client's business, educational, clinical, or research purposes. Insight Assessment can offer only a few points of general advice:

Aggregate vs. Matched-pairs Comparisons. Group comparisons can be analyzed statistically in a matched pairs approach (which associates each individual's post test score with his or her pretest score), or, when the groups are not composed of exactly the same set of individuals, as aggregations of scores. When possible we recommend using the matched pairs approach for pretest post-test comparisons.

Individual Gains: When the same individuals have taken the test at two time points (before and after a treatment designed to train critical thinking skills), one can measure gains by examining difference scores for each individual (Time $_2$ - T_1).

Discarding False Tests: Some tests scores may need to be discarded as uninformative. Some individuals lack sufficient internal motivation to engage a cognitively challenging test with genuine effort (fail to finish the test, select answers randomly), and instead they provide a falsely low test score. Possible indicators of a false test:

1) When your group analyses are performed by Insight Assessment tests with fewer than 60% of the questions answered are dropped from your *group analyses* because they represent incomplete assessments. Scores for these test takers are included in your Excel® file of test takers, however. If you have downloaded this Excel® file to analyze the data at your agency, we recommend dropping these cases as not representative of your test taker group. If there are significant numbers of these cases, consult with Insight Assessment about the difficulty level match with your sample.

2) Very low total scores should be regarded as true scores unless they can be determined to be false scores. These scores indicate that the test taker has very weak critical thinking skills, and these should not be discarded because they represent true scores in your sample. However, any test score that falls in the lowest percentile range when compared to the norm group (0-5th percentile) can be examined as a possible false test. Some possible reasons why very low Total Score might be false include language proficiency problems (contact Insight Assessment for authorized translations of the test) or distractions in the testing center.

3) Critical thinking skills do not deteriorate over short periods of time unless there is an intervening cognitive injury, so the observation of a significant drop in total score from pretest to post test for a given individual is an indicator of a false test at post test. One can examine difference scores from pretest to post test (post test score – pretest score = difference score) and conservatively set a value as indicative of a likely false post test score (any difference score of -3 or more on the 34 point scale or -3.5 or more on the 100 point scale) as evidence of a false test.

4) If there is reason to believe that a given test-taker had serious difficulties reading the language of the test, then that person's test might be eliminated when calculating group results. If the testing environment was affected by an emergency, or by undo noise, commotion, or disturbances which caused test takers to lose time or to be significantly off task, then those tests might reasonably be eliminated as invalid. If a test taker required special needs considerations and these considerations were not afforded, then that person's test might be eliminated when group statistics are being calculated. Test scores of individuals who do not make a true, honest and sustained effort to respond to test questions might reasonably be dropped when calculating group results.

5) The final determination with regard to which tests are or are not to be included when describing group results resides with the client. Best practice suggests determining ahead of time what principles will be used to decide which, if any, tests to exclude.

Categorical Differences: Qualitative evaluations of individuals and groups can be made by examining test scores in relationship to cut scores provided for determining relative strength in overall critical thinking skills (**See the test's User Manual)**. Scale cut scores can be used to determine relative strengths/weaknesses by scale area. Qualitative descriptions of group data can be reported in terms of the percentage of individuals who fall in each area, e.g. 'strong,' 'satisfactory' and 'weak' critical thinking skills. Scale scores can be used in this manner as well. Qualitative improvements in groups can be reported in terms of the change in percentage of individuals in each area or who move across a category, e.g. from a 'moderate/satisfactory' score to a 'strong' score. Categorical gains (e.g. moving from ambivalent to positive on a dispositional inventory, or moving from adequate to strong on a skills test) as described in the test's User Manual are important markers.

Numerical Changes: On skills tests each gain of one additional question correct on the total score represents an important marker of gain in skills, even if the average gain for the group is not statistically significant due to the small sample size. On dispositional inventories a marginal change in the group's numerical average score which remains within the same category (e.g. "ambivalent") may be statistically significant but not otherwise as important as a change group average change that moves from one category to another.

Gains in Relationship to Sample Size: Sample size is an important factor in statistical analysis. Larger gains are required for statistical significance to be attained in smaller sized samples. A group gain of two points is educationally significant for the group overall and likely represents very significant gains in many individuals within the group. If there are fewer than 30 persons in the group, however, statistical tests will report this range of gain as insignificant numerically.

Representativeness: We recommend caution when attempting to generalize from small sample results to assumptions about the population as a whole, unless the sample of test-takers is representative of the larger population. For example, the test results from a sample of 200 students, all of whom have volunteered to be tested, may not be representative of the larger population of students. Similarly, test scores from a sample of freshmen who are residential students may not be representative of the larger population of undergraduates if this larger group includes distance learners, transfer students, and adult part time students.