All Scores Are Not Equal: Evaluating Score Reliability to Improve the Interpretation of Results From Tests of Intelligence in Clinical, Forensic, and School Settings

Gordon E. Taub[1]

[1] University of Central Florida, USA

Correspondence: Gordon E. Taub, University of Central Florida, USA.

Received: April 14, 2022 Accepted: May 10, 2022 Online Published: May 18, 2023

doi:10.5430/irhe.v8n1p18 URL: https://doi.org/10.5430/irhe.v8n1p18

Abstract

Tests of intelligence are administered in clinical, forensic, and school settings. The score most practitioners are familiar with is the Full-Scale Intelligence Quotient (FSIQ). An individual’s FSIQ score is often used for program eligibility and is the most reliable score. In addition to the FSIQ, practitioners often evaluate an examinee’s performance on factor/composite scores. These scores may be used to evaluate an individual’s personal strengths and weaknesses but are less reliable than the FSIQ score. The least reliable scores on an intelligence test are the test’s scaled scores, which are derived from raw scores on individual subtests. It is important for clinicians and practitioners to be aware of the interpretation of scaled scores, but also of their relativly low reliability and higher variability when compared to the FSIQ and factor/composite scores. Because of their lower reliability and variability, clinicians are urged to use extreme caution when interpretating an examinee’s performance on an intelligence test at the scaled score level.

Keywords: psychological testing, assessment, forensic, school psychology, intelligence, scaled scores, FSIQ

1. Introduction

The FSIQ represents an individual’s global intellectual score. It is based on a standard score format, which has an average score of 100 and a standard deviation of 15. An individual who obtains a FSIQ score of 100, scored in the Average range when compared to similar age peers. A score of 115 is in the High-Average range, whereas a score of 130 is in the Gifted or Superior range. Conversely, moving down from the Average FSIQ score of 100, a person with a score of 85 is in the Low-Average range and a FSIQ score of 70 is in the Intellectually Disabled range.

The standard deviation of 15 informs the test user how far an individual’s FSIQ score is from the mean or average of the test. In the example above, the FSIQ scores of 85 and 115 are one standard deviation below and above the test’s mean, respectively. Whereas the FSIQ scores of 70 and 130 are two standard deviations below and above the test mean/average, respectively. The FSIQ score of an examinee is compared to the population and provides an indication of the examinee’s relative standing compared to same age peers. It does not provide an indication of an examinee’s personal strengths and weaknesses.

2. Factor Scores

In addition to the FSIQ score, most contemporary tests of intelligence provide factor or composite scores. Factor scores may inform the examiner of the examinee’s relative strengths and weakness on the areas measured by the test. For example, the Wechsler Intelligence Scale for Children-V (WISC; 2015) provides standard scores on five factors: Verbal Comprehension, Visual Spatial, Fluid Reasoning, Working Memory, and Processing Speed. These five factor scores are reported in a standard score format, identical to the FSIQ, with a mean of 100 and a standard deviation of 15.

3. FSIQ or Factor Score Interpretation

The FSIQ score is used to compare the examinee’s performance on the test to the performance of *others*. Whereas the factor scores may be used to identify the examinee’s *personal* strengths and weaknesses. The decision to use scores that compare the examinee’s scores to the performance of others (FSIQ scores) or to him/herself (factor/composite scores) depends on the purpose of the assessment. For example, if the purpose of the assessment is program eligibility (giftedness/intellectual disability), the FSIQ score (comparing the examinee's performance to others) is the only score that should be interpreted. This assumes, program eligibility is based on the examinee obtaining a specific score. For example, an FSIQ score of 130 may be used as a cuttoff for eligibility in a gifted program or an FSIQ score of 70, may be a cuttoff score to be identified as an individual with an intellectually disablity. (Note, the standard error of measurement is not being discussed herein, but should be applied to any observed score.)

The interpretation of factor/composite scores (comparing the examinee to him/herself) do not impact the decision-making process for gifted/intellectually disabled eligibility. Specifically, a student with an FSIQ score of 131 could qualify for a gifted program. This decision is based on the student's FISQ score compared to the population. If factor scores are examined to identify the student's individual strengths and weaknesses, this analysis would *not* impact the decision-making process. For example, if the student earned a relatively low score on the Visual Spatial factor of the WISC-V, such as 85, but obtained a FSIQ score of 131, the examinee's relatively low score on Visual Spatial factor- would *not* affect the decision-making process. This is because the FSIQ score alone is used to determine program eligibility. The same reasoning applies to an individual being evaluated for an intellectual disability.

In contrast, the factor scores may be examined to identify an individual's *personal* strengths and weaknesses. For example, an individual may have a relatively high score on the WISC-V's factor/composite Working Memory and relatively low score on Fluid Reasoning. This indicates the individual performed better in activities involving Working Memory (holding information in one's mind and transposing it), when compared to Fluid Reasoning (novel problem solving). This form of comparison may be carried out across all factor/composite scores provided by the test. The identification of an individual's relative strengths and weaknesses may be used to assist in program planning or to develop interventions to assist the examinee's personal needs and real-world outcomes.

4. Scaled Scores

An intelligence test is a battery of several individual tests, called subtests. The total number of items answered correctly on each of the subtestsis a test's *raw* score. The raw score from each subtest is converted into a scaled score. A scaled score is a standard score, but on a different scale than the factor scores and the FSIQ score (i.e., mean of 100 and standard deviation of 15). Scaled scores have a mean of 10 and a standard deviation of 3.

Like factor scores, scaled scores may be used to identify an individual's relative strength and weaknesses. The scaled score from each subtest contributes to the calculation of an examinee's performance on the test's factors. For example, on the WISC-V, the Verbal Comprehension factor is a combination of scaled scores from the Similarities subtest and the Vocabulary subtest. Many clinicians will compare an examinee's scaled scores on the Similarities test with the scaled scores on the Vocabulary subtest to identify the examinee's relative strength and weakness on the Verbal Comprehension factor. Because scaled scores are less reliable than factor scores, the interpretation of an examinee's performance at the scaled score level is also less reliable.

5. Score Interpretation and Reliability

Score reliability refers to the consistency of the observed scores and the variability one may see in a re-evaluation (Sattler, 2018). If a score has high reliability, it is considered more stable and more likely to be consistent on re-evaluation. Scores that are less reliable, are more variable and are more likely to deviate on a subsequent evaluation.

When a clinician interprets an FSIQ score, factor score, or scaled score the *stability* or reliability of the score should be kept in mind. Scores that are derived from many tests are most reliable (i.e., FSIQ). Whereas scores from individual subtests (i.e., scaled scores) are less reliable. Reliability is based on a scale of 1.00. This means a test with a reliability coefficient of 1.00 is perfectly reliable. As the reliability of decreases below 1.00, the results become less consistent, and the interpretation of the results is less meaningful/stable.

For example, the overall reliability coefficient of the FSIQ score on the WISC-V is .96. The FSIQ is most reliable score because it is derived by combining more subtest scores than any other factor on an intelligence test. The high reliability of the FSIQ is important because it is often used for decision making purposes.

As clinicians begin to interpret an examinee's performance at the factor/composite level, the reliability decreases. This is easily observed on WISC-V. The reliability of the five factor scores on the WISC-V range from a high of .93 on the Fluid Reasoning factor to a low of .88 on the Processing Speed factor.

When a psychological professional’s evaluation of an examinee’s relative strengths and weakness moves from the *factor score* level to the *scaled score* level, they are making interpretations using relatively unreliable scores. Depending on the examinee’s age, the reliability of the scaled scores is as low as .67 on the WISC-V. This means there is relatively more error surrounding the examinee’s score at the scaled score level, when compared to factor scores or FSIQ. This also means when the examinee is revaluated, psychological professionals should expect to see more variability in the examinee’s performance at the scaled score level. Therefore, the clinician’s interpretation at the scaled score level is less accurate. Because of the lower reliability and variability associated with scaled scores, psychological professionals should use extreme caution when interpreting an examinee’s performance at the scaled score level and avoid this level of interpretation, if possible.

6. Summary

Contemporary tests of intelligence provide psychological professionals with several scores to assist in the interpretation of an examinee’s performance. The FSIQ score is a combination of all individual subtest scores within a test battery, such as the WISC-V. The FSIQ is used to compare the examinee’s score to the performance of others. FSIQ scores are often used for decision-making or eligibility determinations, such as giftedness or intellectual disability.

Scores at the factor level are a combination of scores from individual subtests, but *not all* subtests. Factor scores may be useful when identifying an individual’s personal strengths and weaknesses. Results at the factor score level may be used for program or intervention planning. Next are scaled scores, which are derived from a single subtest. Some clinicians compare the scaled scores that contribute to an individual factor score. This is done to examine narrow strengths and weaknesses at the individual subtest level.

When interpreting the results from the administration of an intelligence test it is important to be aware of the reliability of the scores being evaluated. The FSIQ score is the most reliable score. For example, the reliability of the FSIQ on the WISC-V is .96. Moving further down in reliability are composite or factor scores. The reliability of the factor scores on the WISC-V range from a low of .88 to a high of .93.

Scaled scores are the least reliable score because they are derived from items within a single subtest. The reliability of scaled scores on the WISC-V is as low as .67. Because scaled scores are less reliable, they are also less stable, and less likely to be replicated on reevaluation. This also means the clinician’s interpretation at the scaled score level is less stable and reliable, when compared to the factor scores and the FSIQ score. Therefore, clinicians should use extreme caution when interpreting an individual’s performance at the scaled score level and should avoid scaled score interpretation when possible.

References

Sattler, J. M. (2018). *Assessment of children: Cognitive foundations and applications* (6th ed.). La Mesa, CA: Jerome M.

Wechsler, D. (2014). *Wechsler Intelligence Scale for Children* (5th ed.). Bloomington, MN: Pearson.

Copyrights