Reliability and Validity

Pearson
TalentLens

Score Reliability

The reliability of a test is a measure of the consistency of scores; that is the extent to which two people of the same ability or the same person tested on different occasions will receive the same score. Reliability is expressed as a coefficient which ranges from zero to one. The closer the reliability coefficient is to 1.00, the more reliable the test and the less measurement error there is associated with test scores. **When tests are used in employment contexts, reliability coefficients above .89 are generally considered excellent, .80–.89 good, and .70–.79 adequate. Values below .70 suggest the test may have more limited applicability**. For example, it might be used to provide developmental feedback, but would not be appropriate for making selection or promotion decisions.

A number of methods can be used to estimate test reliability. These include the internal consistency of the test items (e.g., Cronbach's alpha coefficient and split-half reliability), test–retest reliability (the stability of test scores over time) and alternate forms analysis (the consistency of scores across alternate forms of a test).

Validity

Validity refers to the degree to which specific data, research, or theory support the interpretation of test scores (AERA et al., 1999). "Validity is high if a test gives the information the decision maker needs" (Cronbach, 1970). Construct validity is the extent to which the test measures the theoretical construct or trait it is designed to measure. Criterion validity is the extent to which a measure is related to an outcome, e.g. job performance, pass/fail on a course.

There is no set standard for interpretation

- **Below 0.2 – unlikely to be useful**
- **0.2 to 0.35 – useful**
- **0.35 or above – highly effective**

Values are lower than for reliability estimates because:

- Job performance is difficult to predict
- Job performance is poorly measured
- Restriction in range often occurs

Watson Glaser III

Reliability

Internal Consistency: **.82 -.86** (UK and US samples)

Test Retest: In progress currently. Previous studies show correlations of **.73 - .89**.

Alternate Form: **.80** (UK and US samples)

Construct Validity

Studies have demonstrated Watson-Glaser correlates with other cognitive ability measures:

- Correlations with achievement tests ranged from **.39** to **.51**
- Correlations with other reasoning tests ranged from **.47** to **.70**
- Correlations with IQ tests range from **.41** to **.60**
- Moderate correlations with Raven's APM from **.37** to **.53**

Criterion Validity

Watson-Glaser and the Bar Professional Training Course

- In 2011, 1501 students on the course completed Watson-Glaser III.
- Final exam grade was gathered at the end of the course.
- The correlation between the course results and Watson-Glaser was **.51**.
- This research was updated in 2015 using the same method with 998 students
- This data once again confirmed the strong correlation between the test and course performance with a figure of **.55**.

Ravens

Reliability

- APM-III internal consistency: **.73** (International sample of 466 applicants to a UK higher education course in 2015).
- APM-II internal consistency: **.74** to **.84** depending on the sample
- APM internal consistency: **.83** to **.87** depending on the sample
- APM test-retest reliability **.91** (n=243, testing interval of six to eight weeks)

Construct Validity

Years of previous studies on the 36-item APM version (before it was shortened to 23 items) support its construct validity.

- In a sample of 149 college applicants, APM scores correlated **.56** with maths scores on the American College Test (Koenig, Frey, & Detterman, 2007).
- In a study using 104 university students, Frey and Detterman (2004) reported that scores from the APM correlated **.48** with scores on the Scholastic Assessment Test (SAT).

Evidence of construct validity for the current version of the APM is supported by several findings.

- In a subset of 41 respondents from the standardisation sample, the revised APM scores correlated **.54** with scores on the Watson-Glaser (Watson & Glaser, 2006).
- Further, in a sample of N = 276 Raven's APM correlated r = **.51** with the total score on the Advanced Numerical Reasoning Appraisal (ANRA; a cognitive ability assessment measuring quantitative reasoning).

Criterion Validity

Studies using the APM in the past 70 years provide evidence of its criterion-related validity:

- Chan (1996) - In a validation study of assessment

centres, reported that scores on the Raven's Progressive Matrices correlated with ratings of participants on "initiative/creativity" (r = **.28**).

- Gonzalez, Thomas and Vanyukov (2005) - reported a positive relationship between scores on the Raven's APM and performance in decision-making tasks.
- Fay and Frese (2001) - found that APM scores were "consistently and positively associated with an increase of personal initiative over time".

NDIT

Reliability

Internal Consistency: **.76-.78** (UK and US samples)

Construct Validity

NDIT was compared to 'Reasoning for Business Managerial & Graduate—Numerical' (RfB MG Numerical;Cubiks, 2009). Correlation between both tests: **.71**

Criterion Validity

NDIT and Job Performance

- US sample of 104 participants
- Rated their own overall job performance and their performance on the aspects of their jobs that require 'math'.

Total sample: all participants

Those who use Math: reported using math in their jobs at least sometimes, as opposed to rarely or never

Group	*N*	Overall Job Performance	Performance related to math
Total Sample	104	.19	.52**
Those who use Math	71	.32**	.50**

SOSIE

Internal Consistency (UK)

Reliability values range from **.75** (Responsibility and Cautiousness) to **.83** (Dominance) for the Personality Scales, and from **.61** (Conviction) to **.87** (Variety) for the Values Scales.

Internal Consistency (International Sample)

Alphas ranged from **.62** (Conviction) to **.85** (Variety). Two Scales had an alpha <**.70** (Materialism and Conviction). Personality Scales alphas ranged from **.75** (Vigour) to **.81** (Dominance and Sociability). Interpersonal Values Scales alphas ranged from **.70** (Support) to **.81** (Conformity), and Personal Values Scales alphas ranged from **.62** (Conviction) to **.85** (Variety).

Test Retest (UK)

Coefficients ranged from .66 (Responsibility and Sociability) to **.86** (Cautiousness) for the Personality

Scales and from **.53** (Recognition and Orderliness) to **.82** (Conformity) for the Value Scales.

Test Retest (International Sample)

Coefficients ranged from **.78** (Responsibility and Sociability) to **.84** (Dominance and Cautiousness) for the Personality Scales and from **.55** (Conviction) to **.79** (Variety) for the Values Scales