Using tests in the real world

Nils Myszkowski, PhD

Outline

Test administration

Test scoring

Aims of this session

- At the end of this session, you should be able to:
  - Know the basics of test administration
  - Know the basics of test scoring
  - Understand contemporary issues in test administration and scoring

Outline

Test administration

Test scoring

Test administration

- Adequate administration of psychological tests requires preparing:
  - The test environment
  - The test taker
  - The test administrator
  - The test scorer

Test administration

Test administration

Testing environment

Test takers

Test administrators and scorers

Testing environment

- Preparing the testing environment notably requires to eliminate any potential source of distraction.
- To do so, check for:
  - Adequate ventilation
  - Enough space
  - No noise or external stimuli (food, drinks, etc.)
  - Nobody entering the room
  - Providing material (pencils, erasers, etc.)
  - Other people than the test taker and the test administrator

Testing environment

- Of course, group testing has more requirements, notably making sure that test takers are seated in a way that prevents conversations, cheating, etc.

Test administration

Testing environment

Test takers

Test administrators and scorers

Test taker

- The test taker also requires preparation.
- Notably, test administrators should put effort into creating a harmonious relationship and a friendly atmosphere from the beginning of the testing session, even in selection contexts.
- Therefore, it is necessary for test administrators to build rapport, in order to reduce test anxiety.

Test taker

- More than creating a friendly atmosphere, it is also necessary to eliminate any infavorable disposition.
- Notably, test administrators should put effort into engaging interest and cooperation in the testing process, notably by explaining simply the purpose of the test.
- This notably helps having test takers respond openly and honestly during the testing session.

Test sophistication

- Different test takers may be more or less trained in taking a specific kind of test (which is known as **test sophistication**).
- To reduce this source of non-trait variability, test administrators must make sure that they provide explicit instructions and practice items before the test, as it is regularly indicated in the test manuals.
- Failing to provide explicit instructions, even though these instructions may seem obvious for the test administrator sometimes, may invalidate test results.
- Don't take the understanding of the instructions for granted, and remember how it feels when you take an exam and you're not sure what the professor's question is!

Test administration

Testing environment

Test takers

Test administrators and scorers

Test administrators and scorers

- Test administrators should be prepared to the standardized procedures for the administration and scoring of tests.
- These procedures are extensively explained in test manuals.
- Therefore, test administrators should carefully read the manual and be prepared to scrupulously respect these procedures.
- Additionally, test administrators should be ready to build rapport with test takers to reduce test anxiety, and to respond potential questions from the test takers.

Test administrators and scorers

- Individual testing usually requires a lot more preparation that group testing.
- The reason why is because test questions are often presented orally in individual testing, so test takers have to be very familiar with the items and their sequence, and for potential additional rules (timing, starting and stopping points, scoring the responses, etc.).

Outline

Test administration

Test scoring

Test scoring

- Test scores lead to interpretation and these interpretations lead to decisions.
- For this reason, test scoring is a crucial part of the testing procedure.

Scoring rules

- Most of the time, psychologists will "simply" apply the manual's rules for scoring.
- It is important to know them and to be trained in scoring before administrating the test, as it often changes the way you administer the test (making sure the scoring will not require as less as possible of your subjectivity).

Scoring procedures

- Test scoring procedures typically rely either on:
  - The responses of others (the test is then called a **Norm-Referenced Test**)
    - For example, the SAT, GRE, WAIS, WISC, or curved grading systems
  - A fixed criterion decided on during test construction (the test is then called a **Criterion-Referenced Test**)
    - For example, the Driving Test, most exams

Test scoring

Standard scores and percentiles

Score uncertainty

Advanced scoring strategies

Standard scores

- Test interpretation often relies on the use of standard scores.
- The use of standard scores for interpretation is frequent when the aim is to locate an individual in their reference group (for example, in a reference group of children of similar age).
- Standardizing scores also allows to compare the results of different tests (for example, a verbal intelligence test and a numerical intelligence test), even though the raw scores of these different tests were initially on different scales.
- You can think of standardizing as putting every test result "in the same unit", or "on the same scale".

z scores

- z scores have a mean of 0 and a standard deviation of 1.
- For any individual A, a z score of z_A means that the individual has a score that is z_A standard deviations away (can be higher or lower is z_A is positive or negative) from the mean.
  - For example, a z score of $-.04$ means that the individual's score is $.04$ standard deviations **lower** than the mean of his reference group.
- The computation of z scores is achieved using Hull's formula, with the raw score x_A, the mean of the raw scores in the reference group $\bar{x}$, and the standard deviation in the reference group SD_X:

$$z_A = \frac{x_A - \bar{x}}{SD_x}$$

Standard score

- For the purpose of communicability of the results, we can change the z scores to other standard scoring systems, giving them a specific mean and standard deviation.
- For example, in the case of the IQ scores, for historical reasons, a mean of 100 and a standard deviation of 15 are used: $IQ = 15 \times z + 100$.

Standard

- Typical standard score systems (Urbina, 2004):
  - T-scores ($M = 50$, $SD = 10$), used in personality inventories, such as the Minnesota Multiphasic Personality Inventory (MMPI) and the California Psychological Inventory (CPI).
  - Wechsler scale subtest scores ($M = 10$, $SD = 3$), used notably for the Wechler scales
  - Wechsler scale deviation IQs, ($M = 100$, $SD = 15$),used for the summary scores of Wechsler scales and other tests, including many that do not label their scores as “IQs.”
  - College Entrance Examination Board (CEEB) scores ($M = 500$, $SD = 100$), used by the College Board’s SAT as well as by the Educational Testing Service for many of their graduate and professional school admission testing programs, such as the Graduate Record Exam (GRE).

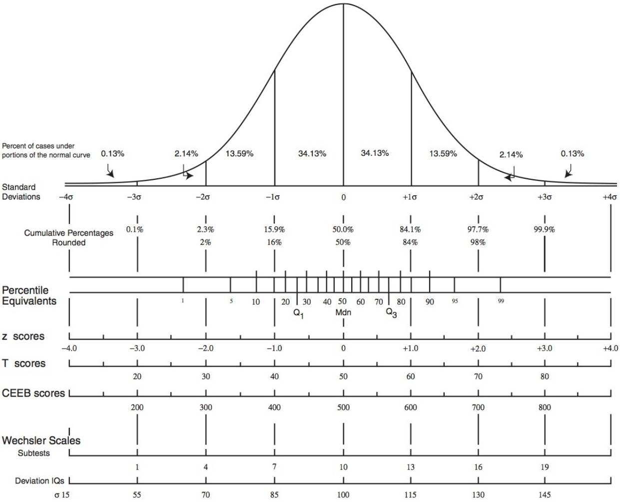
Percentiles

- Under the assumption that scores are distributed normally, to each score is associated a percentile, which represents the area under the curve before the score, which is the percentage of people with an score that is inferior.
- It is of course very useful to intuitively compare individuals with the population.

PACE UNIVERSITY

Test scoring

Standard scores and percentiles



Test scoring

Standard scores and percentiles

Score uncertainty

Advanced scoring strategies

Score uncertainty

- In the reliability section, we have seen that an observed score is only an *estimate* of the true score.
- We have seen that we could estimate the overall reliability of a measure, which is related to how much error the observed scores contain.
- That amount of error was named the **standard error of measurement**.

Standard Error of Measurement

- SE_M is calculated through reliability $\rho_{xx'}$ and population standard deviation σ_x as:

$$SE_M = \sigma_x \sqrt{1 - \rho_{xx'}}$$

- In practice, we will replace population standard deviation by an estimate s and reliability $\rho_{xx'}$ by an estimate $r_{xx'}$.

Score uncertainty

- The standard error of measurement estimates how variable the repeated measures of a person using the same instrument would be around their or her "true" score.
- The standard error of measurement provides an estimate of the expected margin of error of a score.
- Consequently, it can be used to denote that margin of error.

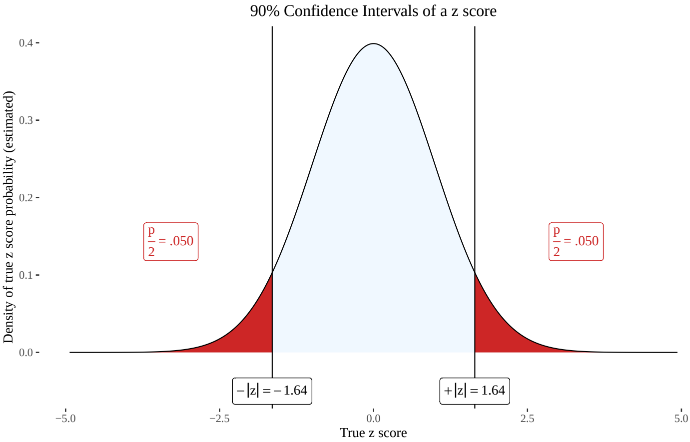
Score uncertainty

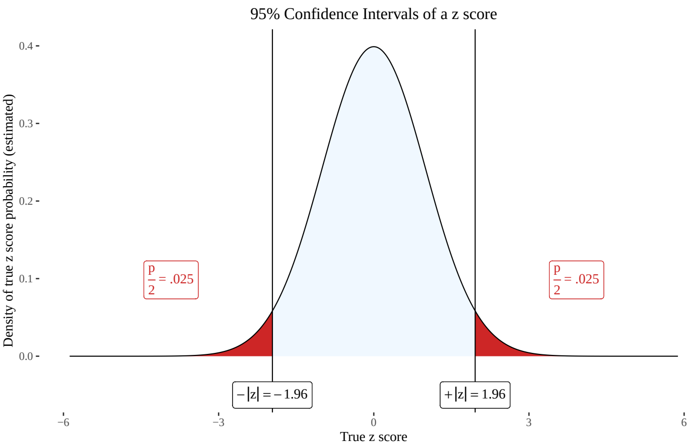
- Alternatively, confidence intervals (CI) may be used.
- They assume a normal distribution of the test scores.
- They are computed so that it is estimated that a certain percentage of the observed scores (the confidence level C) for an individuals would be expected to fall within them.

Score uncertainty

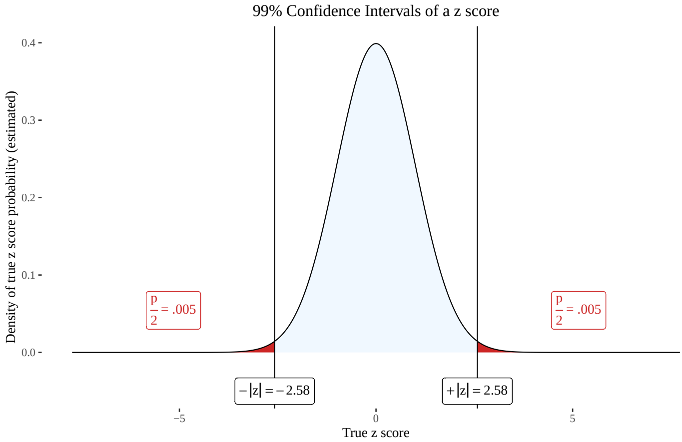
- Confidence intervals are computed by:
  1. Computing the α threshold associated with the confidence level C, through $\alpha = \frac{1-C}{2}$.
  2. Finding the critical $z_{critical}$ value associated with α.
  3. Computing the CI around any x score through:

$$CI_{C\%}\left(x - z_{critical} \times SE_M, x + z_{critical} \times SE_M\right)$$


90% Confidence Intervals of a z score
Density of true z score probability (estimated)
0.4
0.3
0.2
0.1
0.0
p/2 = .050
p/2 = .050
−|z| = −1.64
+|z| = 1.64
−5.0
−2.5
0.0
2.5
5.0
True z score


95% Confidence Intervals of a z score
Density of true z score probability (estimated)
0.4
0.3
0.2
0.1
0.0
p/2 = .025
p/2 = .025
−|z| = −1.96
+|z| = 1.96
−6
−3
0
3
6
True z score

Test scoring
Score uncertainty



Score uncertainty

Confidence Level (C)	$z_{critical}$
99%	2.576
95%	1.96
90%	1.645

Table: Frequently used critical z values

Test scoring

Standard scores and percentiles

Score uncertainty

Advanced scoring strategies

Sum scores

- Most scoring procedures imply summing (or averaging) item scores.
- When doing that, we however sum scores that have a part true score and a part measurement error.
- For that reason, there are strategies to attempt to get rid of (some of) the error part in the score, to have a more "pure" score.

Factor scores

- The main strategy to get rid of error in scoring is to use the estimates of the common part of the items.
- Such estimates are called **factor scores**.

Factor scores

- They are available from statistical techniques that estimate such latent constructs, which are:
  - Exploratory Factor Analysis (EFA)
  - Confirmatory Factor Analysis (CFA)
  - Principal Components Analysis (PCA)
  - Item-Response Theory (IRT)
- There are different methods to compute factor scores from these estimations.

IRT specificities

- Item-Response Theory has some specificities in factor scoring that are often desirable.
  - Scores are directly estimates for the θ_i parameters.
  - Models are usually estimated with the assumption that θ_i follows a Standard Normal $N(0, 1)$ distribution: There is no need for any z score transformation.

IRT specificities

- Item-Response Theory has some specificities in factor scoring that may be desirable.
  - Standard errors vary for each person (which allows to recommend additionnal testing to minimize them).
  - Factor scores can be computed with missing data with minimum bias (missing data is essentially skipped, and thus the only problem resulting from them is an increase in the standard error for the person with missing responses).
  - Missing data can be replaced by plausible values imputation (if there is a need to then compute sum scores, the imputation from IRT modeling is a good choice).

Factor scores

- Be careful, although they sound ideal, the quality of factor scores still depend on the quality of the model. It's also the case for regular sum/average scoring (except they is based on untested axioms).
- Another disadvantage is the low practicality of factor scores in scoring by hand (more or less impossible).
- Finally, factor scores are dependent upon the quality and availability of the data used.