Psychometric Properties & Test scores

Or…

How to Understand Test Scores & Whether a Test is Any Good

Nancy Ryba Panza, Ph.D.
Associate Professor, Psychology Department
California State University, Fullerton

Why bother?

- Why not get to the tests already?
- Because we don't have time to cover all the tests you might see in your cases
- This knowledge applies to all psychological tests & will help you evaluate the quality of the tests used in the work psychologists do for you or for opposing attorneys

This morning…

- Learn about psychometric properties – three major considerations in judging the strength of a test:
  - Reliability
  - Validity
  - Standardization
- Gain an understanding of the types of scores used in most psychological tests:
  - Percentiles
  - Standardized scores

Psychometric Properties

- Aspects of a test that tell you how good it is
- Ways you can judge the strength of a test

- Obviously, if a test is being used to speak to a forensic issue, it must be strong
  - It is the psychologist's responsibility to consider the psychometric properties of the tests they use
  - Must ensure the test is appropriate & strong enough to use for any particular person in any given situation
    - A great test isn't great for everyone or for all situations

Reliability

- Key word: Consistency

- Reliability is the degree to which a test measures something consistently
  - Is someone going to get the same score each time the test is taken?
- A reliable test will produce consistent scores.
  - Think of a scale measuring someone's weight
  - Stop #1 - if you don't have reliability, nothing else about the test matters!

Reliability

- There are many different types of reliability or ways to evaluate the consistency of a test
- Two of most common:
  - Test-retest reliability – does someone get similar scores when taking same test two different times
  - Internal consistency (alpha coefficient) – does someone score consistently on similar items throughout a test
- Results of reliability studies will be in test manual.

Reliability

- All types of reliability are represented quantitatively by a correlation coefficient
- A number that ranges from 0.00 to 1.00
  - The closer to 1.00, the more reliable
- Generally, when using a test to make decisions about an individual, reliability should be >.90
  - There may be exceptions to this rule

Reliability

- Once you know the reliability, you have an idea of how much the test score might change if the test was taken again.
- Reliability tells you how much error is related to a given test score
- Can use this to calculate a confidence interval:
  - A range of scores that represents where the person may score if taken again and again
  - Confidence intervals are less precise, but more accurate ways of reporting test scores
  - This point will be illustrated with the WAIS-IV later

Reliability: In Sum

- Consistency of test scores
- Reliability is a continuum, not a yes or no issue
- Tells how much error you have to consider when interpreting a test score
- Can use to create range of scores (confidence interval) to have better idea of "true" test score

Validity

- Key Word: Accuracy
- A test is valid if it measures what it is supposed to measure and does so accurately.
- This is the most important aspect of a test because if it isn't valid, no conclusions can be drawn from it. It is useless!

Validity

- Validity is also NOT a yes or no issue.
- It is a collective consideration of everything that is known about how a test works.
  - For whom it works
  - Under what conditions it works
  - What conclusions can be drawn
  - What limitations must be considered

Validity

- Validity is a fluid issue that changes over time.
  - Initial validation is done by the test creator.
  - Must show it works right to be published.
- After that time, ongoing research studies and clinical use can provide further evidence for use of the test or limitations for use.
  - Show the test works for additional groups, under other conditions, with different race or age groups, etc.

Validity

- Different types of validity speak to different ways we can feel confident the test works and measures what it intends to measure.
  - Content validity – got the right items in the test
  - Criterion-related validity – test works the same way other similar measures work, can predict outcomes accurately
  - Construct validity – any way can gather evidence to show the test is working

Validity: In Sum

- Validity: Accuracy, Usefulness
- Validity is a collective issue – consider all evidence and make a judgment about how well the test works.
- Psychologist should be aware of validation studies done both prior to and after publication of a test being used.

Standardization

- Most tests in psychology are interpreted by comparing the score someone gets to the scores of others
  - Called norm-referenced tests
  - Different from classroom exams!
- Normative group/Standardization sample
  - Group of people who's scores will be used for comparison
  - Frame of reference for interpreting test scores
- The process of gathering the test scores is called test standardization

Standardization

- Interpretations made from a test are only as good as the standardization.
  - Can't know how well someone does if the group you are comparing his/her scores to is not representative of the people for whom the test is intended.
- So MUST have a representative group
  - If the norms don't represent the population for whom the test was intended, they are useless and all comparisons made will be useless.

Standardization

- Things to consider:
  - Size of group
  - Age range
  - Race/Ethnicity
  - Gender/Sex
  - Geographic representativeness
  - SES
  - When norms were collected (some get dated quickly)
  - Any other factor that could affect the scores

Standardization

- It is the psychologists responsibility to only use a test that is appropriate for the person they are testing.
  - Should not use a test that has not been normed on populations similar to the person being tested because the comparison may not be accurate

Standardization: In Sum

- The process of gathering data to use as a comparison group for test interpretation
  - Norm group, standardization sample
- Must ensure the group provides an appropriate comparison for the person being tested
  - If not, any conclusions made can be tentative, at best

Psychometric Properties

- Three important issues to consider when selecting, and then using, a psychological test.
  - Reliability - consistent test scores
  - Validity - relevant, meaningful test scores
  - Standardization - appropriate comparison group
- The expert using the test must have considered these and is responsible for only using tests that are appropriate for the person and situation.

Understanding Test Scores

Understanding Test Scores

- Three things to understand regarding scores seen in most commonly used psychological tests:
  - Raw Scores
  - Percentiles
  - Standardized Scores

Here's the secret….

- Psychological test interpretation is all about comparisons.
  - How did someone do in comparison to others?
  - Did they score the way most other people do?
  - Did they score much higher than is typical?
  - Did they score much lower than average?

- This works for many kinds of tests…
  - IQ
  - Clinical scales
  - Personality scales

Understanding Test Scores

- Why do we do this?
- Because there are many different tests that measure the same or similar constructs
  - IQ Tests – dozens of good ones out there
    - Differ in number of items, difficulty level, types of items
    - So can't directly compare the actual score someone gets on these tests – 50% right on one test may be VERY different from 50% right on another test
- So we convert the raw scores on tests to other scores that we can use for all tests
  - Like changing to the same language

Understanding Test Scores

- How does it work?
  - Start with the person's raw score (# they got right on the test)
    - 60/100 on an IQ test (60% correct)
  - Convert that score into a universal score that represents how similar or different they scored from others who take the test
    - Compare to norms
      - if average score = 30 – this is really good!
      - If average score = 80 – this is not so good!
    - Change raw score to a score that shows how close or far the person's score is from the average or typical score for their age/gender, etc.

Understanding Test Scores

- There are many different types of scores that raw scores can be changed to in order to make a comparison
  - Percentiles
  - Standardized Scores

Percentiles

- An easy comparison that tells the percent of the population that the person scored above
  - 80th percentile – did better than 80% of people who took the test
- Very commonly used
- Remember: this doesn't tell you their score at all, just how they did in comparison to others
  - NOT a percentage, but a percentile

Percentiles

- So get 60 out of 100 on an IQ test…
  - If this falls at the 25th percentile – not so god
  - If this falls at the 85th percentile – very good
- Get an immediate understanding of how they performed
  - Can easily calculate – but don't have to!
  - There will be conversion charts in test manual using norm group for comparison
    - Just look up raw score obtained & will find percentile
    - Most tests use different charts for age groups, gender, etc.

Percentiles & Standardized Scores

- Percentiles are great but they are limited because they only tell you what "place" someone came in - doesn't tell you how close or far that score is from the other scores

- Alternative comparison: standardized scores

Standardized Scores

- There are many & they all work the same way.
- There is a middle (mean or average) & there is a set amount that describes how much scores vary or how far they are from the average or most typical score (standard deviation)
- Use these two numbers to convert a raw score to a standardized score – one that tells you how a person scored in comparison to others
  - T Score, IQ Score

T Scores

- Used on many personality & clinical scales
  - MMPI-2, CPI, PAI, etc.
- T Scores (Mean = 50, Standard Deviation = 10)
- Any time see a score of 50 = average
  - The higher above 50, the better they did
- Use SD to know how FAR above or below average
  - T score of 70
    - 2 SD above the mean – very high (98th percentile)
  - T score of 40
    - 1 SD below the mean – moderately low (16th percentile)

T Scores

T Score	Percentile	Descriptor
>70	>98th	Very high & very unlikely
70	98th	Clinically significant
60	84th	Borderline
50	50th	Average
40	16th	Below Average
30	2nd	low, maybe clinical
<30	<2nd	Very low & very unlikely

Regardless of the test, any time you see a T score this is the "code" to understanding someone's performance

IQ Scores

- Work the same way as T scores
- IQ score: Mean = 100, SD = 15*
  - So anytime you see an IQ of 100 you know it is average

IQ Score	Percentile	Descriptor
130	98th	Very high, genius
115	84th	Above Average
100	50th	Average
85	16th	Low Average
70	2nd	Very low, Mental Retardation

- *A few IQ tests use a SD of 16 instead

IQ Scores

- This is the "code" for understanding the scores used on IQ tests.
  - Doesn't matter how many items, the difficulty level, what subtests included – because once transformed to IQ scores the scores all mean the same thing
- Will show more when talk about the WAIS-IV

Test Scores: In Sum

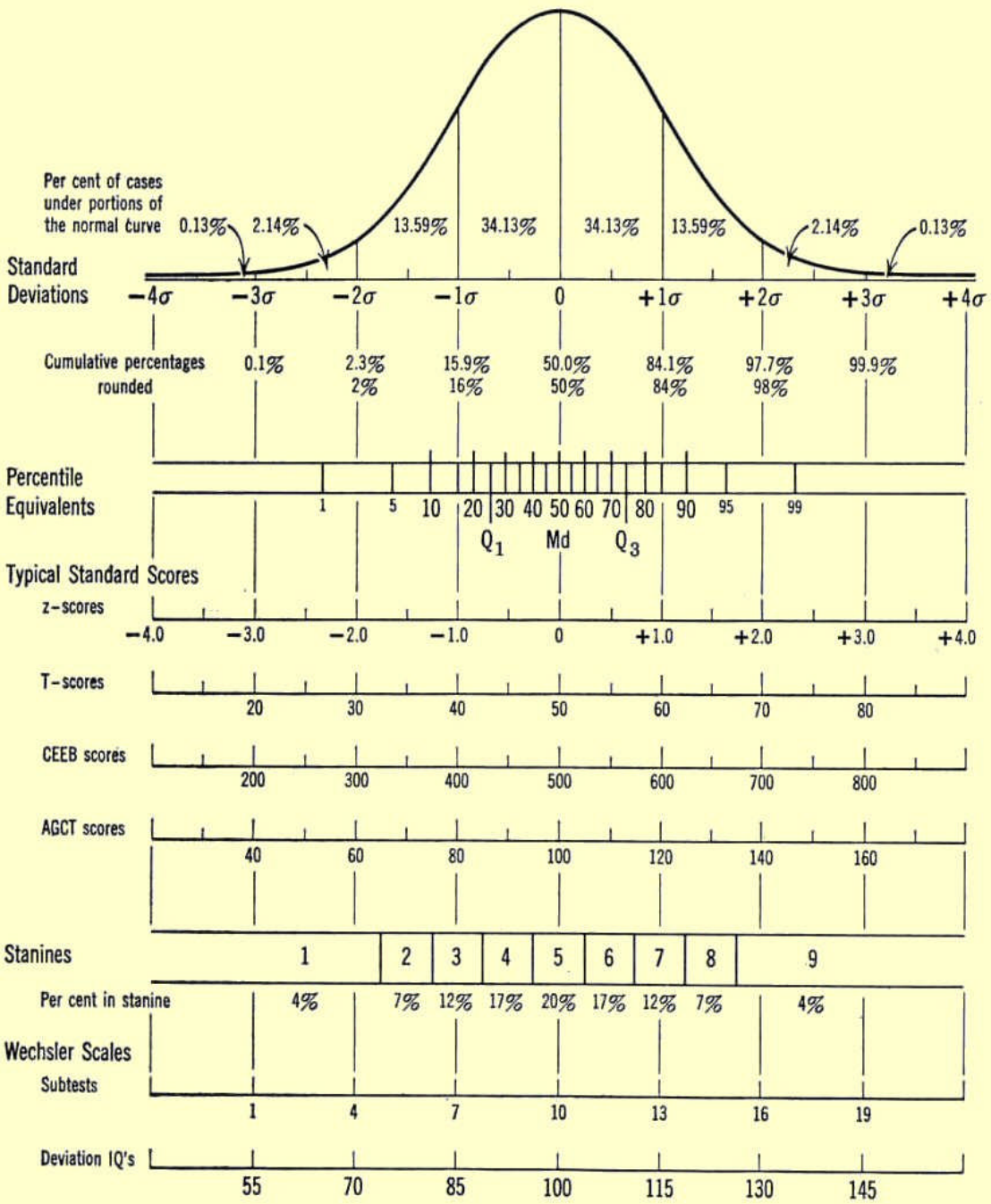
- Raw scores don't mean anything
- Must convert to some type of comparison-based score
  - Percentiles
  - Standardized scores
- Once converted, these scores tell you how close to or far from average a person scored
  - Doesn't matter what test they took, same conversion is used
  - Clinical interpretation may differ from test to test

Test Scores: In Sum

- T Scores & IQ Scores are just 2 types of standardized scores – there are many!
  - Each has a preset Mean & SD (these vary)
  - Each test will have conversion tables for whichever type of score is used

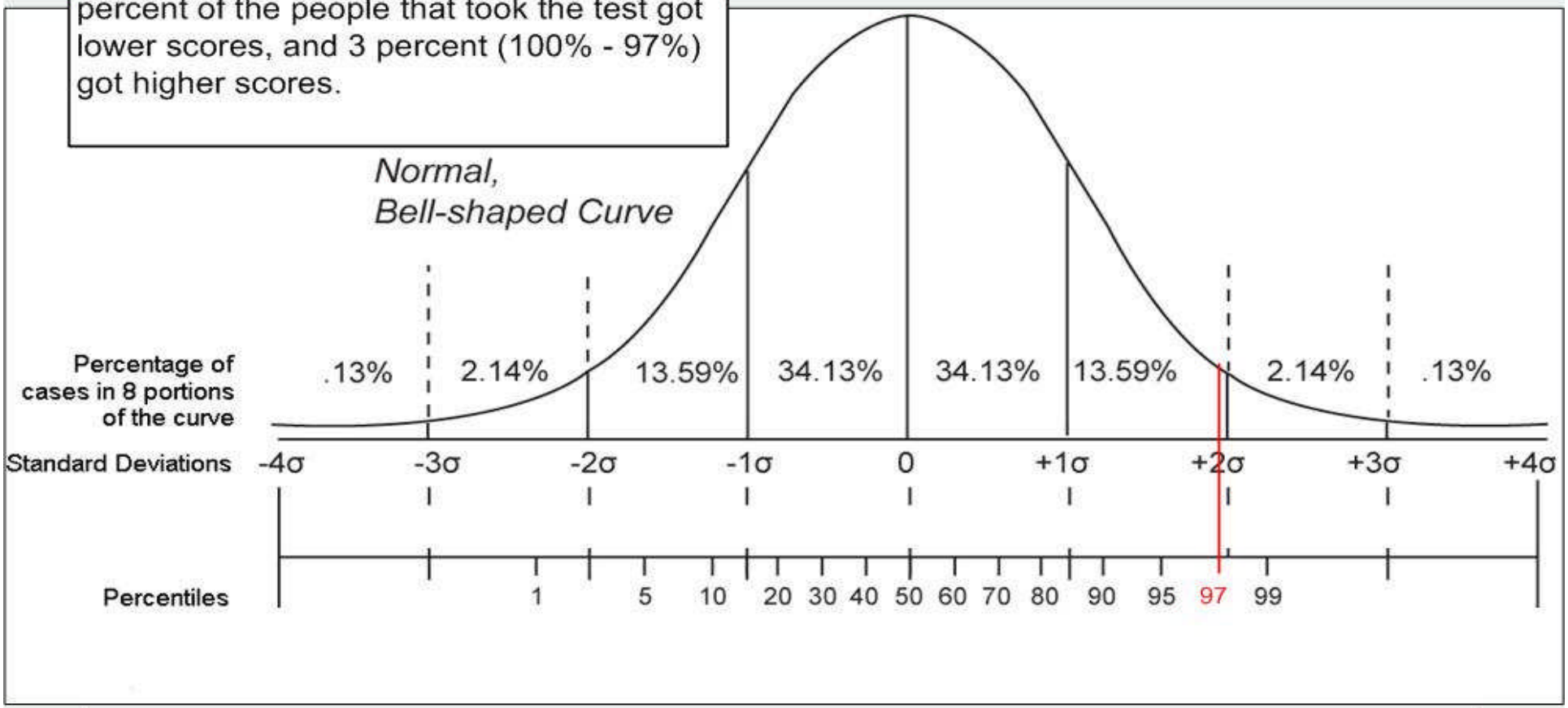
- See graph

NORMS AND UNITS FOR MEASUREMENT



Percentiles

A percentile rank shows the percentage of people that scored above and below a certain score. For example, if a score falls in the 97th percentile, it means that 97 percent of the people that took the test got lower scores, and 3 percent (100% - 97%) got higher scores.



Bell curve labeling was adapted for our purposes. Source of bell curve: http://en.wikipedia.org/wiki/File:PR_and_NCE.gif