Assessment and Testing (Level B)
(3 days of training)

Presented
By

Central Alberta Regional Consortium

Presenter
Dr. Dave Carter

Presented Via Zoom

November 28, 29 and 30, 2023

9.00 – 3.30 MDT (Alberta time)

Participants should print this handout package single sided - we will use it throughout the three days.
Note that attendance at all three days is required to be awarded the certificate of completion.

Welcome!

1. Some ideas and thoughts…

- The first purpose of these sessions is to give all participants an in-depth knowledge of the principles, practices, issues, strengths, weaknesses, uses and abuses of "Level B" tests.
- These sessions are **NOT** about how to administer and score one or two common Level B tests… that is just training to do one thing. These sessions are **in-depth** and provide the knowledge and skills to select, administer, score and interpret a range of Level B instruments. Including ones not yet published.
- We will take the time necessary to give each person a detailed, practical, and "comfortable" knowledge of measurement statistics, standard scores, test administration and interpretation.
- Key tests include (but are not limited to): WJPB-Achievement, WIAT, KeyMath Diagnostic Arithmetic Test, Alberta Diagnostic, KTEA, PPVT, Canada QUIET, and others…

Some provocative quotes…

"Complex psycho educational testing is "…the science of the strange behaviour of children in strange situations with strange adults for the briefest periods of time."

Bonfenbrenner, 1977

- *"Substantial evidence indicates that the same treatment goals and teaching strategies are adopted regardless of the category of mild disability."*

  Rechsley & Ysseldyke, 1995

- *"…Dave, the head of standardized achievement testing has been lopped off… it just refuses to fall!"*

  Anonymous, 1991

- *"The least initial deviation from the truth is multiplied later a thousand-fold."*

  Aristotle, c. 1200 B.C.

- *"…(in effect) little errors in the beginning lead to serious consequences in the end."*

  Acquinas, c. 1200 A.D.

- *"The beat of a butterfly's wing in Shanghai effects the weather in Beijing."*

  Anonymous, ancient

2. The "Big Picture" of testing…. history… how did we get here?

- From Binet to Wechsler
- From Alpha – Beta
- From Vineland – SIB-R
- SES, gender, race… the confounding variables…
- Why is it important (essential) that persons administering, scoring and interpreting Level B tests meet the criteria?

4. What are Levels A, B, and C? What are the rules?

Level A tests: These are tests which can adequately be administered, scored and interpreted with the aid of the manual, a familiarity with the client population, orientation to the kind of setting within which the testing is done, and a general knowledge of measurement principles and of the limitations of test interpretations. This category includes most interest inventories, group or individual, and multiple-choice tests that employ a simple metric as the main avenue of interpretation (e.g. occupational clusters).

Level B tests: **These are tests that require specific training for administration, scoring, and interpretation. These tests are more complex than Level A tests and require sophisticated understanding of psychometric principles, the traits being measured, and client population and clinical uses involved in the setting within which the testing is done. This category would generally include most individual or group tests of achievement or interest, screening inventories and personnel tests.**

Level C tests: These tests require advanced (graduate level) training for interpretation in the specific professional field to which the tests apply… some of these tests may also require this level of training for competent administration and scoring. These tests are more complex than Level A and B tests… this group would generally include any aptitude or language or personality or clinical diagnostic test, group or individual.

Now…prepare to take the plunge into measurement theory and other such esoteric and wonderful things….

5. Some key terms (the "yin and yang" of testing….)

Standardized *versus* non-standardized

Norm referenced *versus* criterion referenced

Group *versus* individual

Screening *versus* diagnostic

Pre *versus* post

6. **So… what is so "normal" about the "Gaussian Curve"???? Settle down and relax… this takes a while but is well worth the knowing.**

   · Shoes sizes…. Wear them if they fit…

   · Pine seeds… but only if they are random enough!

A KEY CONCEPT:

The normal curve is approximated in sufficiently large and random samples.

… only approximated….. that's all!

Standardized tests should be based on:

- Statistically large samples (how big is big enough)?
- Random samples (consider "depth" of randomization)?
- Samples from where?
- Samples from when?

Here is a little example to think about….

Mary Jones got a percentile rank score of 50 on the XYZ test of reading… so her score is "half way" up in her grade seven class… she is in the middle…. (percentiles later in course)

BUT……

The XYZ test of reading was published in 1995!
It was standardized on 300 children in Florida, Alabama and California!
15% of the population in the samples were Hispanic!
20% of the population in the samples were Black!
The test was anchor-normed on a similar test standardized in Alaska in 1978!

QUESTION – what is wrong with this? Answer….

So… what about Canadian tests?

Does a maple leaf on the golden arches make it a "Canadian Burger"?

US *versus* Canadian samples… some essential information:

- American tests dominate the Canadian market (either directly or indirectly)….
- That is not necessarily "bad" as long as you are knowledgeable and careful…
- Anchor test norming….
- IQ and SES north and south of the border… are Canadians "smarter"… eh?

8. Meanwhile, back at the "normal curve."

Get your pencils ready…

The **four** properties of the normal curve:

- Mean ( )
- Mode ( )
- Median ( )
- Symmetricity

NOTE: To be a "normal curve" all four of the properties that define a normal curve must be present….

Dave will draw for you some examples of normal and non-normal curves… make copies of his sketches.

9. **Now for the mathematics… we will walk carefully and inexorably through the manual calculation of the fundamental statistic required for Level B testing…..** ***why is he doing this to us?????***

Use a grid - here…

Remember the Carter definition of "standard deviation":

The degree to which the average score varies from average… on average!

The formula for variance is:

The formula for standard deviation is:

Now… remember the relationship between variance and standard deviation:

Variance = s.d. squared

s.d. = square root of the variance

Now… we need to go back to the "shape" of the normal distribution.

Remember it is a normal distribution when the _____,

_____, and _____ are in the same place and it is

_____ about the mean.

Now….. Dave will explain the "mantra of the normal curve…"

34 – 14 – 2 Cha!!!

Draw a normal curve in the space below and label the mean, as well as the standard deviations…. Dave will show you how….. and we will add some common score transformations…

T score
z score
Standard score
Deviation IQ score

All you need is the mean and standard deviation for each…

- **T score. mean = 50; s.d. = 10 (mnemonic is _____)**
- **z score. mean = 0; s.d. = 1 (mnemonic is _____)**
- **Standard Score. Mean = 100; s.d. = 15 (no mnemonic)**
- **DIQ Score. Mean = 100; s.d. = 15 (unless you are an OLD Stanford-Binet)**

10. An exercise.

What is the T score equivalent of a z score of +2?

What is the DIQ equivalent of a T score of 30?

11. Meanwhile we need to turn our thoughts to "purity."

- The raw score (aka the "observed score")
- Why the raw score is "pure"
- Why we often neglect the raw score
- Why it is the "genesis" of all other scores.
- A practical and true illustration of the importance of the raw score: "…galloping brain rot"!

Remember – everything beyond the raw score is "transformed" and is therefore not 100% pure….. **remember the meaning of the word "frugal"!!!**

13. The most easily misunderstood score….

- percentile rank scores (Gauss on his side)
- the use of "cut scores"

REVIEW OF PART ONE:

- The normal curve is approximated in sufficiently large and random samples.
- It has an inherent shape and is defined by the fact that the mean, mode and median all occur in the same place and that it is symmetric about it's own mean.
- The purest score is that on which all converted scores rests – the raw score.
- Standard score conversions like the T, z, DIQ and standard score can easily be converted if you know the mean and standard deviations of each.

Now – onward!!!

14. Reliability and Validity.

Even when you have the numbers… it is still about these two!

Reliability versus Validity

Reliability is "the degree to which a test can be counted on over time."

- Sharon at the mall.

- Rubber tape measures.

Validity is "the degree to which a test does what it claims to be able to do."

- Eat lots of nourishing food…. depends on which way you read the data…

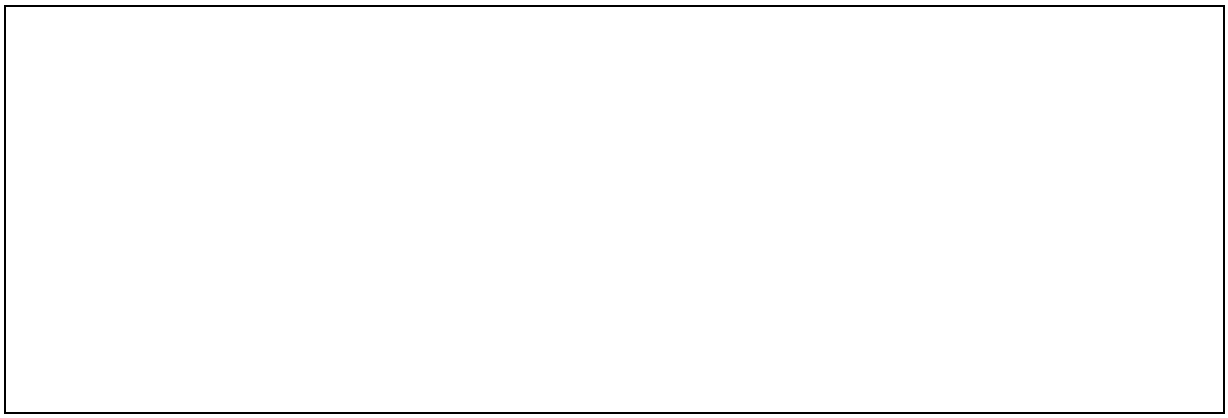
The key relationship between reliability and validity and the biggest key to understanding measurement.

High reliability does not necessarily equal high validity

BUT

High validity almost always equals high reliability.

Now for a lovely little artistic rendering….



The dynamic relationship between reliability and validity is vital to a full understanding of measurement. Validity always leads the way… control it and you are doing something important. Reliability by itself means little!

You will come back to the "V R Railroad" over and over again if you are involved in education measurement in a thoughtful and informed manner… it is vital to ALL forms of educational measurement…

15. The Standard Error of Measurement (SEm)

- The difference between a "true" or "universal" score and an "observed" score.

- Todd's question… "who knows the true score?"

- All true scores are made up of the observed score plus or minus some "error." LIKE THIS:

X obs. = X true +/- error!

So… how about a test with "zero error"?

- The very best tests have small (but never zero) error… but not every time!!!

- Note: you can to a large degree control test error by obtaining high validity… get high validity and the reliability tends "to take care of itself."

- Comments about two kinds of validity later…

SEm is the distribution of scores on repeated theoretical administrations of the same test.

Illustration of SEm:

Note that the more valid a test is, generally the more reliable it is. If a test is valid and reliable it should have a lower SEm all other things being equal…

Every observed score (written as Xi) is made up of the true score plus or minus some error (we said that already).

The SEm is a reflection of how reliable a test is (hopefully because it is valid).

Commercial tests calculate and report (in the manual) a measure of the degree to which observed scores can be relied upon.

They do this by calculating the SEm for observed scores on the test.

CAUTION: the SEm is usually reported based on the overall averages of all calculated SEms… that can lead to some problems for children "at the ends of the normal curve."

16. Every observed score = true score +/- some error.

Like this (remember): **X true = X obs +/- error**

The amount of this error can be **ESTIMATED** but never known for sure…

Why can it only be estimated?

Commercial tests use the term "Confidence Interval" (CI).

A Confidence Interval may be reported at one or more "levels of confidence."

You do not calculate them… they are given to you in the manual.

The test may refer to the Confidence Interval OR SEm… but usually it is referred to as the Confidence Interval….

Commonly: + / - 1 SEm is appx. The 68% CI

+ / - 2 SEm is appx. The 95/96% CI

The confidence interval is simply the standard deviation of the distribution of scores estimated to occur on repeated administrations of the test. The standard deviation is calculated in the same way as we did previously – but it is re-named the SEm as opposed to the s.d. and is the basis for the reported CI…. Simple… actually it is quite easy in practice…

Example 1

- Mary took the XYZ test of reading comprehension. She got an observed raw score of 29 which yielded a standard score of 80.
- The SEm of the test is reported as 10

- We are therefore 68% "sure" that her true score is somewhere between

  _____ and _____.

- We are therefore 96% "sure" that her true score is somewhere between

  _____ and _____.

Example 2

- John took the ABC test of arithmetic calculation. He got an observed raw score of 48 which yielded a standard score of 70.

- The SEm of the test is reported as 8.

- We are therefore 68% "sure" that his true score is somewhere between

  _____ and _____.

- We are therefore 96% "sure" that his true score is somewhere between

  _____ and _____.

Let's turn it around a bit… Juanita was given a standard score of 105 on a standardized test of reading. It is noted that the upper and lower limits of the 68% CI are "110 and 100." Question – what is the SEm of the test?

Here is another one… Sam was given a standard score of 95 on a standardized test of spelling. It is noted that the upper and lower limits of the 96% CI are 105 and 85. What is the SEm of the test (careful!!!).

QUESTION: "Why should we bother to calculate confidence intervals – what is their purpose?"

17. Coefficient of Determination.

Very spiffy technical phrase for one of the MOST USEFUL little statistical calculations in the book….

This can be used for both "reliability" and "correlation"

It goes like this:

R squared = Coefficient of Determination

Explanation:

The coin flip mnemonic!

Coefficient of Determination = R squared…

0.8 …….. _____ Coefficient of Determination

0.7 …….. _____

0.6 …….. _____

0.9 …….. _____

1.0 …….. _____ ??!!??

18. Now a "detour to a naughty score." (some cautionary notes)

The "non-equal interval score" – GEq (and a better way!!!)

GEq are highly contrived…

a. non-equal interval

b. Not taken across the school year…. usually only once and the rest is presumed….

c. Suggests a level of specificity that is not possible.

d. gain score problems:

Raw Score	GEq.
17	2.1
18	2.1
19	2.4
20	2.7

Mary's raw score increased from 17 to 19 after three months of remediation.
Her Grade Equivalent score rose from 2.1 to 2.4… only three months gain.

Jim's raw score increased from 18 – 20 after three months of remediation.
His Grade Equivalent score rose from 2.1 – 2.7… a six month gain !!!???

It would be better to use percentiles……

Or – use GRADE RANGES – early, mid, high….

Incidentally you can "work backwards" in a test manual to get a range…..

19. Issues of test selection.

- Define the purpose of the test FIRST.
- Look at validity – especially "Content" and especially not "Face"
- Get technical information – not just "…this is the newest and best thing since sliced bread!"
- Check external reports and the reports of other users(careful with Google)!

- Look carefully at the technical information about the sample and sampling procedure used to generate the norms for the test:
  - How many?
  - Who?
  - Where?
  - When?
  - Comparability to your subjects.
- Consider "other factors"
  - Cost of the test.

- Cost of the protocols.
- Speed of administration.
- Ease/accuracy of scoring.
- Computer support (not for analysis though!)
- "Grade" versus "Age" norms – or both?
- Check to see if your district has any policy/practice guidelines About which level B tests can be used by whom…

The mechanics of test administration:

· Reading the manual…. Why and what?

· Practicing tests on a "no stakes" child

· Fully understand the basal and ceiling rules!

The purposes of basals

The purposes of ceilings

Use the basal and ceiling rules carefully and always note if there has been any variations. Watch out… basal and ceiling rules change from test to test and often from subtest to subtest!

The single biggest contributors to test error:

· Incorrect use of basal and ceilings

· Adding and subtraction errors

· Mistakes moving around inside the "tables"

- Question – do you have to generate all of the available scores (from percentiles to NCEs to RPIs)???

- Retention of the protocol – legal issues

- Ownership of test protocols and reports… watch our what you write down… you may be looking at it under oath!

- Test security – why tests should not be allowed to float into the public domain… especially but not restricted to Level C tests.

Appendix 1
Key Diagnostic Questions

1. Are the questions, issues, concerns that prompted the referral valid?
2. Have possible "organic aetiologies" been considered – most especially have hearing and visual acuity actually been measured and if necessary ameliorated?
3. What is the child's current level of social adjustment in school? Is behaviour or maladaption a problem?
4. Is the child's behaviour (academic and social) variable across domains and time?
5. Have the results of assessments been shared as appropriate with other involved professionals? Are people talking and sharing?
6. Is there any compelling need for further assessment of any kind?
7. If the child is to be labelled in any way – will this pragmatically improve their chances of success?
8. Are assessment results going to contribute to a meaningful/measurable IPP/IEP?
9. Will the IPP seek to address the problem of "generalizability" of any gains and are Goals and Objective clearly stated and measurable.
10. Are Goals and Objectives only being measured by standardized tests? If so that is a problem… beware of regression effects and the inherent lack of accuracy in Level B tests.
11. Is there appropriate consideration for further Level B and Level C assessment later? This is especially true where Level C results have resulted in a child being "classified." When will this classification be revisited? Can things change? Could we be wrong? What are the implications of not thoroughly reviewing classificatio