PSY 554
Tests and Measures
Jim Graham

Scope of Practice
- You will not be administering and interpreting psychological batteries; interpretation is the domain of the above licensed psychometrician.
- As a school counselor, your work will often be focused on serving the best interest of your students. You will be doing this in an environment where the use and misuse of measurement is rampant. It will be necessary for you to understand measurement in order to help your charges.
- Mental health and school counselors will be working in areas where people will expect you to understand measurement. Working with Interdisciplinary teams means that you will be at a disadvantage if you don’t know measurement. Others may expect you to be the expert.
- Researchers – in order to study something, you must measure it.
- You can use assessment for:
  - Identify assess client problems
  - Conceptualize and define the problem
  - Select and implement the effective treatment
  - Evaluate the counseling

Association for Assessment in Counseling and Education
- http://www.theaaceonline.com/
- http://www.theaaceonline.com/resources.htm
- Position Statement on High Stakes Testing
- Standards for Qualifications of Test Users
- Standards for Multicultural Assessment
- Standards for Assessment in Mental Health Counseling
- Standards for Assessment in School Counseling

History of Testing
- Sir Francis Galton
  - (1822-1911, half-cousin to Darwin, true polymath, excelled in many fields)
- Scientific meteorology
- Weather maps
- Method of classifying fingerprints for forensics
- Silent dog whistle
- Perfect cup of tea
- Beauty map of Britain
- Responsible for correlations and seen as the founder of psychometrics.

Intro
- Goal of course:
  - To become familiar with the basic psychometrics necessary for understanding and evaluating assessment procedures
  - Not to administer and interpret tests (557)
- Psychological tests are “objective” and standardized measures of a sample of behavior;
  - “Behaviors” can include intelligence, ability, personality, mood, thoughts, and other characteristics (relationships, past experiences, etc.)
- Tests are systematic procedures for observing behavior and describing it with the aid of numerical scales or fixed categories.

Tell me some reasons it is important to study history.
- Much of his work focused on the heritability of human ability (nature vs. nurture was his as well)
- Used sensory discrimination and reaction time tests to measure intelligence (which is heavier, etc.)
- Determined that human ability is largely hereditary, much of his work was used as the basis for eugenics.
- Galton saw the “average man” as mediocre.

**Eugenics** – improvement of human hereditary traits through different interventions (e.g. selective breeding, sterilization)

**Historic applications:**
- Forced sterilization of individuals with mental retardation; compulsory euthanasia and racial cleansing in Nazi Germany
- The Oregon Board of Eugenics, later renamed the Board of Social Protection, existed until 1983, with the last forcible sterilization occurring in 1981
- Modern applications:
  - Modern eugenic prenatal testing and screening, genetic counseling, birth control, in vitro fertilization, marriage counseling

**Cattell**

- James McKeen Cattell (1860-1944)
- 1st psych prof in the US (University of Pennsylvania)
- Developed idea of “mental tests”, which were largely focused on sensation and perception
- Mental Tests and Measures (1890)
  - Dynamometer pressure (strength of hand squeeze)
  - Sensation areas
  - Pressure causing pain
  - Least noticeable difference in weight

**Binet**

- Alfred Binet (1857-1911)
- In 1895 he stated that intelligence tests were too sensory oriented, headed a French commission on mental retardation
- 1905 – The Binet-Simon scale; 30 items, normed with a sample of 50; measured judgment, comprehension, and reasoning
- 1908 – revised the instrument to include a ratio of mental age to chronological age, called IQ (intelligence quotient)
- 1916 Stanford-Binet (adapted by Terman at Stanford, of “termite” fame; standardized on an American sample of 1000 children and 400 adults

**Terman**

- Lewis Terman (1877-1956)
- Study of gifted children in 1920’s, continued today
  - They were not weak and sickly social misfits, but in fact were generally taller, in better health, better developed physically, and better adapted socially than other children.
- Member of Human Betterment Foundation, a eugenics group involved in the promotion and enforcement of compulsory sterilization laws in California
- Terman administered English-language tests to Spanish-speakers and non-educated blacks to provide scientific proof of the intellectual superiority of whites

**Yerkes**

- Robert Yerkes (1876-1956)
- During WWI, the Army needed tests that could be administered on-mass, rather than individually, used multiple-choice format
  - Army Alpha – used for routine testing
  - Army Beta – used for French or non-English speaking recruits (or those who failed the alpha)
  - Army Individual - re-used administered test for people who failed the Beta
  - [http://www.officialusarmy/armyptsptests](http://www.officialusarmy/armyptsptests)
  - People used from 1A to 4E; scores below 3 were not suitable for officer training
- Findings of research:
  - Average US mental age was 15 (top score of morale)
  - Differences between ethnic origins (Western and Northern Europeans were highest; Slavic and darker southern were lowest (Pole – 16.77)
  - Used to argue for borders immigration laws
  - Later analysis showed this was explained by the number of years they have spent in the US
  - Average score of black men was 56.4, well below white average.

**Sensation areas**

- Counseling, birth control, in vitro fertilization, marriage counseling.
**Testing Ethics**

- Each professional group has their own ethics code:
  - APA: Section 9
  - ACA: Section F
  - ASCA: A.9
- Each of the codes are similar in what they expect.

**Misc. History**

- **Frank Parsons** – “The Father of Guidance” 3-step career counseling: Understand the person, understand the world of work, match the person to the appropriate work
- **Spearman (general intelligence) vs Thurstone (7 primary factors)** What is intelligence?
- After WWI, there was an interest in IDing recruits who are not emotionally capable of serving – **Woodworth’s Personal Data Sheet** (1917) – who will be susceptible to “Shell-Shock”
- 1921 – Rorschach Inkblot Test – Projective, based on psychoanalytic theory

**Misc. History**

- **Strong** – 1927 – Interest Blank (Career)
- **1923** – Stanford Achievement Test – Measures of knowledge and performance in multiple academic subjects
- **1939** – Mental Measurements Yearbook
- **1940’s** – MMPI (Hathaway and McKinley) – Items were empirically selected and tied to criterion, rather than because they appeared to measure different facets

**Misc. History**

- **60s and 70s** – LOTS of testing in schools – instruments begin to be scrutinized for ethnic bias
- **Minimum competency testing** – there should be a minimum amount of knowledge needed to get a high school diploma
- **As for counseling**
- **Modern statistical methods have changed the landscape of testing dramatically:**
  - Item response theory has allowed for the development of computer adaptive tests
  - More sophisticated methods of assessing bias

**Misc. History**

- **1940’s** – MMPI (Hathaway and McKinley) –
- **1939** – Mental Measurements Yearbook
- **1923** – Stanford Achievement Test Measures
- **1927** – Interest Blank (Career)
- **1921** – Rorschach Inkblot Test – Projective, based on psychoanalytic theory

**Misc. History**

- You are bound to behave ethically, and to monitor others for ethical use of tests and measures.
- **Scope of Practice** – Not all tests are within every profession’s scope of practice
  - **Testing companies attempt to sell tests to only those qualified to use them** – Most use a 3-tiered system:
    - A – Books (anyone can use them)
    - B – Brown ADD scales, some self-report, most career inventories, many achievement tests ( Masters level counselors)
    - C – WISC, MMPI, Rorschach (psychologists)
- **Competence** – Only use measures and tests with which you have competence

**Misc. History**

- Consent procedures for testing are generally the same as for counseling
- **Avoid labeling individuals based on a single test score.**
  - A single test score is full of error
  - Nearly all professional organizations explicitly forbid high-stakes decisions being made on the basis of a single test score
- Consider disability, culture, education, etc. when using tests
Assessment security —
- Don’t release raw test data to non-qualified others.
- Maintain the security of tests!

What is a Variable?

Scales of Measurement
- Nominal (aka categorical)
  - A score is not an amount, rather a discrete category
  - The categories are not ordered
- Ordinal (aka ranked)
  - The discrete categories are placed in order
  - The distance between the places are not known

Scales of Measurement
- Interval (aka continuous)
  - The distance between the ordered values are equal (the distance between 2 and 3 is the same as the distance between 3 and 4 or 4 and 5)
  - There is no “true” zero value (true zero indicates the absence of a variable
- Ratio
  - Same as interval data, but 0 indicates the absence of the variable.

Central Tendency
- Measures of central tendency attempt to characterize a set of data with a single number.
Mode

- The most frequently occurring value.
- Could have more than one mode
- Pros:
  - Can be used with all scales of measurement.
- Cons:
  - Not useful for further computations
  - Does not take all of the scores into account

Median

- Midpoint of the distribution: 50% of cases lie above the median, 50% of cases lie below the median
- Cannot be used with categorical data (because there is no order)

To calculate:

- Put all the scores in rank order
- If N is odd, the median is the middle score
- If N is even, the median is the average of the two middle scores

- Pros:
  - Not affected by outliers
  - Considers all scores

- Cons:
  - Does not consider the value of all scores (only order)
  - Not useful in further computations

Mean

- The average of all scores in the data
- Can only be used with interval or ratio data

\[
\bar{X} = \frac{\sum X}{n}
\]

- Pros:
  - Considers the actual value of all scores
  - Very useful for further computation
- Cons:
  - Very affected by outliers

Choosing Measures of CT

- Partly dictated by the scale of measurement you are using.
- You should select the measure that best characterizes your data
The most frequently occurring value is 3.

The middle value is 3.

The sum of x is 27/9 is 3.

Outliers affect only the mean.

Extreme outliers greatly reduce the utility of the mean.

Calculate the mode, median, and mean of this data.
Variability

- Measures of central tendency tell us about a typical observation.
- Measures of variability tell us about how spread out a distribution is.

\[ X \]

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
</tbody>
</table>

\[ \text{Mode} = 1, 5 \]
\[ \text{Median} = 3 \]
\[ \text{Mean} = 3 \]

Which measure of central tendency best represents the data?

Variability

- Range = Highest score – Lowest score
  - Only considers two scores
  - Heavily influenced by outliers

- On average, how much does each score differ from the mean?
- What is the average deviation from the mean?

\[ \bar{X} \]

\[ \sum X = 20 \]
\[ N = 5 \]
\[ \bar{X} = 4 \]

- First, calculate the mean of \( X \)

\[ X \]

\[ 3 \]
\[ 4 \]
\[ 5 \]
\[ 6 \]

\[ 3 \]
\[ 4 \]
\[ 5 \]
\[ 6 \]

\[ \sum X = 20 \]
\[ N = 5 \]
\[ \bar{X} = 4 \]

- Subtract the mean from each \( x \) score – this is a **deviation score**.
Subtract the mean from each score – this is a deviation score.

\[
\begin{array}{c|c|c}
X & X - \bar{X} & (X - \bar{X})^2 \\
\hline
2 & 2 - 4 = -2 & 4 \\
3 & 3 - 4 = -1 & 1 \\
4 & 4 - 4 = 0 & 0 \\
5 & 5 - 4 = 1 & 1 \\
6 & 6 - 4 = 2 & 4 \\
\hline
\sum X = 20 & \sum (X - \bar{X}) = 0 & \sum (X - \bar{X})^2 = 10 \\
N = 5 & \bar{X} = 4 & \\
\end{array}
\]

- The sum and average of the deviation scores will always be zero.
- What can we do to solve this problem?

Square each of the deviations from the mean.
- The sum of the squared deviations from the mean is called the sum of squares.

\[
\begin{array}{c|c|c}
X & X - \bar{X} & (X - \bar{X})^2 \\
\hline
2 & -2 & 4 \\
3 & -1 & 1 \\
4 & 0 & 0 \\
5 & 1 & 1 \\
6 & 2 & 4 \\
\hline
\sum X = 20 & \sum (X - \bar{X}) = 0 & \sum (X - \bar{X})^2 = 10 \\
N = 5 & \bar{X} = 4 & \\
\end{array}
\]

The average sum of squares is called the variance (\(S_X^2\)).
- The variance is the average squared deviation from the mean.

\[
S_X^2 = \frac{\sum (X - \bar{X})^2}{N}
\]

\[
S_X^2 = \frac{10}{5}
\]

\[
S_X^2 = 2
\]

To put the variance back into the same metric as the mean, take the square root.
- This is called the standard deviation.
**Standard Deviation**

- Large standard deviations indicate data that are widely spread out around the mean.
- Small standard deviations indicate data that are closely clustered around the mean.

<table>
<thead>
<tr>
<th>Mode</th>
<th>Median</th>
<th>Mean</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
</tbody>
</table>

- What is the standard deviation of a constant?

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
</tbody>
</table>

- What would a distribution with a $s_x = 10$ look like?
- $s_x = -1.15$?

**Standard Deviation**

- The formula for variance and standard deviation estimates calculated from population data are different from those calculated from sample data.

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Population</th>
<th>Sample (estimated population)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Variance</td>
<td>$\sigma^2_x = \frac{\sum(x - \bar{x})^2}{N}$</td>
<td>$s^2_x = \frac{\sum(x - \bar{x})^2}{n - 1}$</td>
</tr>
<tr>
<td>Standard Deviation</td>
<td>$\sigma_x = \sqrt{\frac{\sum(x - \bar{x})^2}{N}}$</td>
<td>$s_x = \sqrt{\frac{\sum(x - \bar{x})^2}{n - 1}}$</td>
</tr>
</tbody>
</table>

- The population formulae divide by $N$, while the sample formulae divide by $n-1$.

**Standard Deviation**

- The sample mean is an unbiased estimator. It might be high, it might be low, but there is no reason to believe it will be biased in one direction over another.
- Because the standard deviation uses the sample mean (not the population mean), it will tend to underestimate the population standard deviation.
**Standard Deviation**
- The standard deviation can tell us how good of a summary score the mean is.
- The mean is a least squares estimate.

**Correlation**
- In correlational research, we are interested in examining the relationship between two or more variables.
- To examine the relationship between two variables, it is necessary that you obtain measures of both variables from each individual.
- The relationship between the two variables can be represented visually as a scatterplot.
Often, the results are further summarized by plotting a straight line through the points.

We can see that, in general, higher levels of relationship satisfaction is related to lower medical expenses.

Is this a small or a large relationship?

How does this relationship compare to the relationship between exercise and medical expenses?

A correlation (often referred to as “r”) is a number ranging from -1 to +1.

The sign indicates the direction (or “type”) of relationship between the two variables.

The absolute value indicates the strength of the relationship between the two variables.

The correlation coefficient (often preferred to as “r”) is a number ranging from -1 to +1.

The sign indicates the direction (or “type”) of relationship between the two variables.

The absolute value indicates the strength of the relationship between the two variables.

\[ r_{xy} = -0.30 \]

\[ r_{xy} = +r_{xy} = - \]
Correlation Coefficient

Correlation Coefficient

Correlation Coefficient

Correlation Coefficient

Correlation Coefficient

Pearson’s r

The Pearson product moment correlation coefficient is a correlation between two interval or ratio scaled variables.

Pearson’s r is calculated by determining whether the ways in which individuals differ from one another on one variable are reproduced in another variable.

Correlations are a measure of linear relationships.
Always examine your data for curvilinear relationships.

A correlation is:
- A number ranging from -1 to +1
- The closer to 1 the absolute value of the correlation, the stronger the relationship.
- The sign indicates the direction of the relationship.
- A squared correlation indicates:
  - The % of variance (individual differences) that is shared by the two variables
Pearson’s $r$ assumes:
- Data are measured on an interval or ratio scale.
- Observations are paired.
- The relationship between the variables is linear.

\[ r_{xy} = \frac{S_{XY}}{\sqrt{S_{X}^2} \sqrt{S_{Y}^2}} \]

\[ \frac{\sum (X - \bar{X})(Y - \bar{Y})}{n - 1} \]

* For a population

\[ \frac{\sum (X - \bar{X})(Y - \bar{Y})}{N} \]

* Inferring from a sample to a population

Classical Test Theory
Many of the constructs we are interested in as psychologists cannot be measured directly:
- Intelligence
- Mood
- Personality
- Attitudes

As a result, we attempt to operationalize the constructs of interest into a measurable format

Latent Variable (latent construct of interest) → Measured Variable (observed variable)
- IQ Score
- Score on a self-report scale
- MMPI-2 Score
- Reaction time on an implicit association test

Classical Test Theory
\[ X = T + E \]
An observed score is made up of the latent trait of interest (the true score) and factors other than the latent trait of interest (error)
Classical Test Theory

An observed score (or measured) is made up of the latent trait of interest and factors other than the latent trait of interest (error).

We never know how much error is influencing a single measured item score.

We can attempt to better estimate the latent trait of interest by not relying on a single observed score:
- Use multiple items
- Administer the test multiple times

Process of Test Construction

1. Identify the purpose for which the test scores will be used.
   - Passionate love: "A state of intense longing for union with another. Reciprocated love ... is associated with fulfillment and ecstasy. Unrequited love ... with emptiness, anxiety, or despair" (Hatfield & Walster, 1978, p. 9)
   - To be used as a research tool to measure the level of passionate love experienced by an individual in a particular relationship

2. Identify behaviors that represent the construct or define the domain (theoretical work, review research)

3a. Prepare a set of test specifications –
   - How many items will cover each of the domains?
3b. Create a set of item specifications:
   - Type of response scale (T/F, multiple choice, Likert)
   - Number of response choices (2 to 10)
   - Direction of items (forward and reversed scored)

Process of Test Construction

Cognitive
- Intrusive thoughts about or preoccupation with partner
- Idealization of other or of the relationship
- Desire to know the other and to be known

Emotional
- Attraction to other, especially sexual attraction
- Positive feelings when things go well
- Negative feelings when things go poorly
- Longing for reciprocation
- Desire for complete, permanent union
- Physiological arousal

Behavioral
- Try and determine the other's feelings
- Studying the other person
- Service to the other
- Maintaining physical closeness

Item Types

Forced choice

- is constantly on my mind
- Agree / Disagree

Likert
- Sexually compatibility is necessary for a successful relationship.
  - Strongly Disagree
  - Disagree
  - Neutral
  - Agree
  - Strongly Agree

Semantic differential
- Describe your relationship

- Exciting
  - 1 2 3 4 5 6 7 8 9
  - Boring

Likert-like
- Sometimes I feel I can't control my thoughts; they are obsessively on
  - Not at all
  - 1 2 3 4 5
  - Not at all
  - Definitely
  - In-between
Process of Test Construction

- 4. Create a pool of items from your test and item specifications
  - Emotional (negative feelings if things go poorly):
    - “I would feel deep despair if ___ left me.”
    - “I'd get jealous if I thought ___ were falling in love with someone else”
    - “If I were separated from ___ for a long time, I would feel intensely lonely”
    - “An existence without ___ would be dark and dismal”

Crocker and Algina’s (1986) suggestions for item construction

- Put statements or questions in present tense
- Avoid statements that can have more than one interpretation
- Avoid statements that are likely to be endorsed by almost everyone or almost no-one
- Try to have an equal number of positively and negatively worded items
- Keep statements short (<20 words)
- Each statement should be a good grammatical sentence
- Statements containing universals (all, always, none, and never) introduce ambiguity and should be avoided; likewise avoid indefinite qualifiers (only, just, merely, many, etc.)
- Use simple sentences (avoid “if” or “because” clauses)
- Use vocabulary that can be understood easily by the respondents
- Avoid the use of negatives (not, none, never)

Process of Test Construction

- 5a. Have the items reviewed by colleagues
  - Grammar, readability, offensiveness or appearance of bias, appropriateness of items to specifications
- 5b. Revise items based on feedback
- 6. Preliminary item tryouts
  - Have some people take the test, observe behavior during the test
  - Interview the test takers about ambiguous items, etc.

Process of Test Construction

- 4. Create a pool of items from your test and item specifications
  - Cognitive (idealization): “For me, ___ is the perfect romantic partner”
  - Emotional (physiological arousal): “Sometimes my body trembles with excitement at the sight of ___”
  - Behavioral (studying): “I take delight in studying the movement and angles of ___’s body.”

- Create a range of items across difficulty levels
  - Ceiling effect – People get all of them right
  - Floor effect – People get all of them wrong

- 7. Large-scale field test
- 8. Determine statistical properties of items and test scores, eliminate inadequate items
- 9. Conduct reliability and validity studies on final form of test
- 10. Develop norms and guidelines for administration, scoring, and interpretation of the test scores
**Item Scoring**

- Typically, we’re not interested in an examinee’s response to a single question, rather, we use multiple items.
- Scales of measurement – categorical, ranked, interval, ratio.
- Scoring schemes –
  - Sum of items (arbitrary scale, minimal computation).
  - Average of items (scale of subscale same as item scale, or % correct if 0,1; Different scales are not comparable).
  - Standard scores (used as part of norm-referenced tests) – Have a set Mean and SD.

---

**Test Scores**

- Norm-referenced vs. Criterion referenced.

---

**Norm-referenced tests:** Compare the examinee’s performance to the performance of others (IQ, personality, Iowa, etc.).

- Important to clearly define the norm group
  - Representative (Census-matched sample)
  - Current (Flynn Effect)
  - Sufficiently large sample size.
- Scores typically reported as standard scores or percentile ranks.

---

**Criterion-referenced tests:** Compare the examinee’s performance to a pre-determined level of proficiency on the criterion of interest (driver’s tests, MSP/HSPE, psych licensure).

- Content domain (criterion) must be clearly defined.
- Often use cut-off scores or divide scores into performance categories.

---

**Z - Scores**

**Effects of Constants**

- Applying additive and multiplicative constants to a set of data has a known effect.

<table>
<thead>
<tr>
<th>X</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>4</td>
</tr>
<tr>
<td>5</td>
</tr>
<tr>
<td>6</td>
</tr>
</tbody>
</table>

\[
\bar{X} = 4 \\
S_X = 1.41
\]
## Effects of Constants

- Applying additive and multiplicative constants to a set of data has a known effect.

<table>
<thead>
<tr>
<th>X</th>
<th>X + 1</th>
<th>X + 10</th>
<th>X - 3.1</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>3</td>
<td>12</td>
<td>-1</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>13</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td>14</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>6</td>
<td>15</td>
<td>2</td>
</tr>
<tr>
<td>6</td>
<td>7</td>
<td>16</td>
<td>3</td>
</tr>
</tbody>
</table>

\[ \bar{X} = 4 \]
\[ S_X = 1.41 \]
Effects of Constants

- Applying additive and multiplicative constants to a set of data has a known effect.

<table>
<thead>
<tr>
<th>X</th>
<th>X+1</th>
<th>X+2</th>
<th>X+3</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>3</td>
<td>12</td>
<td>-1</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>13</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td>14</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>6</td>
<td>15</td>
<td>2</td>
</tr>
<tr>
<td>6</td>
<td>7</td>
<td>16</td>
<td>3</td>
</tr>
</tbody>
</table>

\[ \sum X = 4 \quad \sum (X+1) = 5 \quad \sum (X+2) = 14 \quad \sum (X+3) = 14 \]

The mean changes by a direct function of the constant:
New Mean = Old Mean + Constant

Effects of Constants

- Applying additive and multiplicative constants to a set of data has a known effect.

<table>
<thead>
<tr>
<th>X</th>
<th>X*2</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td>4</td>
<td>8</td>
</tr>
<tr>
<td>5</td>
<td>10</td>
</tr>
<tr>
<td>6</td>
<td>12</td>
</tr>
</tbody>
</table>

\[ \sum X = 4 \quad \sum (X*2) = 8 \]

The mean changes by a direct function of the constant:
New Mean = Old Mean + Constant

Effects of Constants

- Applying additive and multiplicative constants to a set of data has a known effect.

<table>
<thead>
<tr>
<th>X</th>
<th>X+10</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>12</td>
</tr>
<tr>
<td>3</td>
<td>13</td>
</tr>
<tr>
<td>4</td>
<td>14</td>
</tr>
<tr>
<td>5</td>
<td>15</td>
</tr>
<tr>
<td>6</td>
<td>16</td>
</tr>
</tbody>
</table>

\[ \sum X = 4 \quad \sum (X+10) = 5 \]

The standard deviation does not change as a function of the additive constant:
New Std. Dev. = Old Std. Dev.
## Effects of Constants

- Applying additive and multiplicative constants to a set of data has a known effect.

<table>
<thead>
<tr>
<th>X</th>
<th>$X^2$</th>
<th>$X\times 10$</th>
<th>$X/2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>4</td>
<td>20</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>6</td>
<td>30</td>
<td>1.5</td>
</tr>
<tr>
<td>4</td>
<td>8</td>
<td>40</td>
<td>2</td>
</tr>
<tr>
<td>5</td>
<td>10</td>
<td>50</td>
<td>2.5</td>
</tr>
<tr>
<td>6</td>
<td>12</td>
<td>60</td>
<td>3</td>
</tr>
</tbody>
</table>

$\bar{X} = 4$  
$S_x = 1.41$  
$\bar{X}/2 = 2$  
$S_{x/2} = 2.82$  

The mean changes as a direct function of the multiplicative constant: 
New Mean = Old Mean * Constant

The standard deviation changes as a direct function of the multiplicative constant: 
New Std. Dev. = Old Std. Dev. * Constant

- By using additive constants, we can change the mean of a set of data without affecting the dispersion.
- By using multiplicative constants, we can change the mean and standard deviation of a set of data.

2 inches <> 6 centimeters

- The problem is that these measures are not on the same scale.
- They must be converted to the same scale.

1 inch = 2.54 centimeters
2 inches <-> 6 centimeters
- The problem is that these measures are not on the same scale.
- They must be converted to the same scale.
  1 inch = 2.54 centimeters
  2 inches = 5.08 centimeters < 6 centimeters

**Standard Scores**
- Standard scores are used to place scores into a metric that allows those scores to be more easily compared to other scores.
- Each type of standard score has a given mean and standard deviation.

<table>
<thead>
<tr>
<th>Standard Score</th>
<th>μ</th>
<th>σ</th>
</tr>
</thead>
<tbody>
<tr>
<td>IQ Score</td>
<td>100</td>
<td>15</td>
</tr>
<tr>
<td>T-Score</td>
<td>50</td>
<td>10</td>
</tr>
<tr>
<td>ETS score</td>
<td>500</td>
<td>100</td>
</tr>
<tr>
<td>z-score</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

**Z-score**
- What type of constant can we apply to the data to set the mean equal to zero?

<table>
<thead>
<tr>
<th>X</th>
<th>X - μ</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>.2</td>
</tr>
<tr>
<td>3</td>
<td>-1</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>2</td>
</tr>
</tbody>
</table>

X = 4
S_x = 1.41

**Z-score**
- We could go through and apply the constant to every individual score, and then calculate the mean and standard deviation of the transformed scores.
- However, we also know what the new mean and standard deviation will be without calculating the scores.

<table>
<thead>
<tr>
<th>X</th>
<th>X - μ</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>2 - 4 = -.2</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
</tr>
</tbody>
</table>

X = 4
S_x = 1.41

**Z-score**
- What type of constant can we apply to the transformed data (X - μ) to set the standard deviation to 1?

<table>
<thead>
<tr>
<th>X</th>
<th>X - μ</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>.2</td>
</tr>
<tr>
<td>3</td>
<td>-1</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>2</td>
</tr>
</tbody>
</table>

X = 4
S_x = 1.41

The new mean is equal to the old mean (4) plus the constant (-4)

**Z-score**
- Standard deviations are not affected by additive constants.
### Z-score

What type of constant can we apply to the transformed data \((X - \bar{X})\) to set the standard deviation to 1?

<table>
<thead>
<tr>
<th>X</th>
<th>(X - \bar{X})</th>
<th>((X - \bar{X}) / S_X)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>.2</td>
<td>-1.42</td>
</tr>
<tr>
<td>3</td>
<td>-.1</td>
<td>-.70</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>.70</td>
</tr>
<tr>
<td>6</td>
<td>2</td>
<td>1.42</td>
</tr>
</tbody>
</table>

\[ \bar{X} = 4 \]
\[ S_X = 1.41 \]

The new mean is equal to the old mean (0) times the constant \((1/1.41)\).

The new standard deviation is equal to the old standard deviation \((1.41)\) times the constant \((1/1.41)\).

These are now z-scores, which indicate the number of standard deviations above or below the mean a score falls.

\[ z_x = \frac{(X - \bar{X})}{S_X} \]

Which is the better score:
- SAT Verbal of 550
- WAIS Verbal IQ of 110
- **Common standard scores:**
  - IQ (M= 100, SD=15)
  - T (mean = 50, SD=10)
  - Z (m=0, SD=1)
  - Stanine (standard nine – Mean=5, SD=2)
  - College board (m=500, SD=100)
  - Wechsler subtest (M=10, sd=3)
  - ACT (m=20, sd=5)

- **Normal Distribution**
  - Normal Distribution
    - Mean = Median = Mode
    - Symmetrical
    - Asymptotic to the x-axis
  - Standard Normal Distribution
    - Normal distribution of z-scores

- **Variables that are normally distributed can be converted into percentile ranks.**

- **A percentile rank** is the percentage of cases that a single score is higher than.
  - For example, someone with an IQ of 100 scored higher than 50% of their same-age peers

- **Reliability**
  - In psychology, we are often faced with the prospect of trying to measure constructs that cannot be directly observed (e.g., happiness, love, paranoia).
  - The actual scores on questionnaires attempting to measure these constructs are not treated as direct measures, rather they are also influenced by factors other than the construct of interest.
Classical Test Theory

\[ X = T + E \]

An observed (or measured) score is made up of the latent trait of interest (the true score) and factors other than the latent trait of interest (error).

Reliability

- **Reliability**: the degree to which scores are consistent and do not contain error.
- **Reliability coefficients** characterize the percentage of variance in a measured score that is variance in the true score on the latent trait of interest

\[ r_{xx} = \frac{S^2_{true}}{S^2_{measured}} \]

Most reliability coefficients are calculated as correlations, but interpreted as squared correlations.

Test-Retest Reliability

- If a perfect measure is measuring a stable trait, scores should not differ from administration to administration.
- Any changes in scores from one administration to the next cannot be due to changes in the true score, and must therefore be due to error.

If \( r_{xx} = .8 \), then 20% of the variance in the observed scores are due to error.

Because a variable should be perfectly correlated with itself, the bar for reliability coefficients is much higher than for correlations.

- .7 is often considered the cut-off for "acceptable" reliability
- .8 is actually a better benchmark
- .9 or greater is what most people want, and is a good benchmark when measures will be used to make high stakes decisions.

Test-Retest Reliability

- If a perfect measure is measuring a stable trait, scores should not differ from administration to administration.
- Any changes in scores from one administration to the next cannot be due to changes in the true score, and must therefore be due to error.
Test-Retest Reliability

- Assessing test-retest reliability:
  - Administer a measure to a sample.
  - Wait a pre-specified amount of time (2 weeks, 1 month, etc.)
  - Re-administer the measure to the original sample.
  - Calculate the correlation between the two scores.
  - This coefficient is not useful if the construct being measured is not stable.

Inter-Rater Reliability

- Assessing Inter-Rater reliability:
  - Obtain a sample of behavior from a group of participants.
  - Have two raters independently code all of the behavior.
  - Calculate the correlation between the scores of rater 1 and rater 2.
  - There are other more complex ways of estimating inter-rater reliability (for more than 2 raters, when all raters don’t code all behaviors, etc.).

Internal Consistency

- If it is assumed that each item is measuring the same construct, measured scores would not differ from item to item in a perfect measure.
- Any changes in scores from one item to the next cannot be due to changes in the true score, and must therefore be due to error.
- This family of coefficients is often used for surveys or questionnaires where multiple items are used to assess the same trait.

Split-Half Reliability

- Assessing Split-Half reliability:
  - Obtain responses to a survey from a group of participants.
  - Randomly split the test in half (or even/odd).
  - Score the two halves of the test separately
  - Calculate the correlation between the two scores.

Cronbach’s Alpha

- On a test with more than 2 items, there are multiple ways to split the test; the split-half coefficient would therefore be different depending on how the test was split.
- Cronbach’s coefficient alpha is the average of all possible split-half reliabilities.
- This is the most commonly used measure of reliability, and is very easy to compute using SPSS.
Reliability

- Reliability is a property of scores, not measures.
- A measure is not reliable or unreliable.
- Reliability is influenced by the measure.
- Reliability is also influenced by characteristics unique to the sample and setting.
- Always calculate reliability for the data in hand!!!

While we can’t determine the exact amount of error in any individual score, we can estimate it:

\[ \sigma^2_T + \sigma^2_E = \sigma^2_x \]

\[ \sigma_{xx} = \frac{\sigma^2_T}{\sigma^2_X} \]

- Solve for the standard deviation of error, or the standard error of measurement (SEM):

\[ \sigma_E = \sigma_x * \text{SQRT}(1 - \rho_{xx}) \]

- Because any observed score is due, in part to error, any score on one administration of a test is not likely to be 100% accurate.
- If random errors of measurement are normally distributed, we would expect 95% of Confidence intervals would contain the true score.

Calculating Cronbach’s alpha in SPSS:

- Make sure that each of the items are scored in the same direction (to reverse score, \( x_i = (k + 1) - x \), where \( k \) is the highest possible response.
- Analyze => Scale => Reliability analysis
- Move the items to the left

The Standard Error of the Measure is:

\[ \sigma_E = \sigma_x * \text{SQRT}(1 - \rho_{xx}) \]

- Because error is assumed to be random, the average effect of error on scores is 0, and the effect of error is assumed to be normally distributed.
- 95% of the time, error will impact a score no more than 1.96 SEM units.
- A 95% CI can be constructed about an individual test score by +/- 1.96 Standard error units.

Other Z-score equivalents:
- 99% CI - 2.33
- 99.5% CI - 2.58
- 95% CI - 1.96
- 90% CI - 1.64
- 80% CI - 1.28

Confidence Intervals

- Due to error, a single score is a poor estimate of one’s ability.
- To take into account the effect of error, the SEM can be used to construct a confidence interval about the test score.

\[ C.I_{95} = x \pm 1.96 \times \text{SEM} \]

95% of confidence intervals should contain an individual’s true score.
- This does NOT mean that a confidence interval is 95% likely to contain a true score.
- Reliability is a property of scores, not measures.
- Example: Measure of verbal ability; Chinese versus American
- Using norms/SEM from other samples doesn’t necessarily fit!
- Other statistics:
  - Alpha if item deleted – Estimated alpha if the item was removed.
  - Item-total correlation – All should be positive, correlation between the item and the sum of the remaining items.
Validity of Scores

- **Reliability** considers the amount of error in measured scores.
- **Validity** considers whether the scores are actually measuring what they are supposed to be measuring.
- Validity supports the types of inferences that are to be drawn from test scores.

Reliability and Validity

- Reliability is a pre-requisite of validity
- Using unreliable measures will deflate correlations using those measures.

Content validation

- **Content validation**: Drawing inferences from test scores to a larger domain of similar items. Often used in achievement tests.
- Define the performance domain of the test
  - Should some domains be given more weight?
  - e.g. – Objectives for 1st grade math: Add any two positive numbers whose sum is 18 or less
  - Subtract any two whole numbers (each less than 20) whose difference is a positive value
- Have independent experts judge whether the items cover the domain of interest
  - Provide a structured framework for matching items to the performance domain

Content Validity

- Collect and summarize data from the matching process
  - % items matched to objectives
  - Correlations between importance and number of items
  - % of objectives not assessed by any item on test
  - **Index of item-objective congruence** (used if each item should only match one objective – aggregates expert’s ratings from +1 to -1)

Content Validity

- **Duplicate construction**: The best way to do this is to have two separate teams develop tests using the same set of specifications;
  - The error determined by comparing the two scores should approach the split-half reliability
- **Bias** – look to see if cultural, racial, ethnic, or gender bias could influence things; e.g., if domain is to test math skills, story problems (which also test English-language skill), are biased against non-native English speakers
- **Face validity**: whether something looks like it measures what it is supposed to
  - Content validity is NOT face validity
  - Sometimes face validity is desirable, sometimes it is not
Criterion Validity

- **Criterion-related validation**: Drawing inferences from test scores to performance on a real behavioral variable of practical importance
- Differentiated into:
  - Predictive validity (criterion is future performance) and
  - Concurrent validity (criterion data and test scores obtained at the same time)

When the criterion data are available, obtain a measure of performance on the criterion

- Determine the strength of the relationship between the criterion and the test score
- **Reliability**: the product of the square root of the reliability coefficients forms the upper limit of the criterion validity (the criterion and predictor can’t be correlated more than allowed by the true scores)
- **Validity coefficient**: correlation
- **Coefficient of Determination**: squared validity coefficient
- **Expectancy tables**: used when you have a categorical criterion – used to determine cut-off points (depends on cost of false positives and negatives)

<table>
<thead>
<tr>
<th>Score on Test</th>
<th>0-9</th>
<th>10-19</th>
<th>20-29</th>
<th>30-39</th>
</tr>
</thead>
<tbody>
<tr>
<td>No re-offense in 5 years</td>
<td>10</td>
<td>9</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>Re-offense within 5 years</td>
<td>0</td>
<td>2</td>
<td>9</td>
<td>10</td>
</tr>
</tbody>
</table>

- **Expectancy table**: Recidivism risk scale:

Cut-off of 19/20 minimizes most false positives and negatives, but what if the consequence of a false positive meant that an innocent person was killed?

Construct Validity

- **Construct validation**: Drawing inferences from test scores to performances that can be grouped under the label of a specific psychological construct
- **Psychological construct**: a product of informed scientific imagination, e.g. “intelligence”, “creativity”, “Personality”
  - Must be operationally defined
  - There must be theoretical links between the construct and other measures or real world criteria

1. Formulate hypotheses about how those who differ on the construct will differ in regards to demographic characteristics, performance criteria, or other measures; These should be based on theory
2. Select instruments to measure the other pieces
3. Gather empirical data
4. Evaluate hypotheses, rule out rival theories and alternative explanations
   1. Correlate the measures
   2. Look at how scores differentiate between groups (MMPI, DIS)
   3. Factor Analysis: examines item intercorrelations; are factors as theorized?
Construct Validity

- **Multi-trait multi-method matrix**

  1 | 2 | 3 | 4
---|---|---|---
1) MMPI Depression | .89 |
2) MMPI Anxiety | .89 | .91 |
3) Beck Depression | .76 | .71 | .60 | .93 |
4) Beck Anxiety | .33 | 1.00 |

- **Convergent validity** is assessed by examining the mono-trait/hetero-method correlations
  - They should be as high as possible, approaching reliability if at all possible

- **Discriminant validity**: The scores shouldn't be correlated with unrelated constructs
  - High hetero-trait/mono-method correlations would suggest error resulting from the method of measurement

- **Reliability coefficients**: On the diagonal, should be high

---

Construct Validity

- **Multi-trait multi-method matrix**

  1 | 2 | 3 | 4
---|---|---|---
1) MMPI Depression | .89 |
2) MMPI Anxiety | .89 | .91 |
3) Beck Depression | .76 | .71 | .60 | .93 |
4) Beck Anxiety | .33 | 1.00 |

- **Reliability coefficients**: On the diagonal, should be high

- **Discriminant validity**: The scores shouldn't be correlated with unrelated constructs
  - High hetero-trait/mono-method correlations would suggest error resulting from the method of measurement

- **Ideally**:
  \[ \text{Reliability} = \text{M-Trait/H-Method} > \text{H-Trait/M-Method} = \text{H-Trait/H-Method} \]

- **Realistically**:
  \[ \text{Reliability} > \text{M-Trait/H-Method} > \text{H-Trait/M-Method} > \text{H-Trait/H-Method} \]
**Reading a Multi-Trait Multi-Method Matrix**

1. Check the reliability coefficients. We want these as high as possible.
2. Check the convergent validity (MT-HM) coefficients. We want these as close to the reliability coefficients as possible.
3. Check the discriminant validity coefficients – these should be lower than the convergent validity coefficients.
4. Check the HT-MM coefficients – the lowest of these has the least shared method variance.

**High Stakes Testing in Education**

- **No Child Left Behind**
  - Proposed in 2001, signed in 2002, re-funded in 2007
  - Focused on theory of *standards-based education reform*
    - Set the standard of what a student should be able to do – raise the bar for all students
    - Not norm-based (how students do compared to one another),
    - “Outcomes-Based Education”
    - Bi-partisan law, Ed Kennedy and Bush
    - Bush promoted it with improvements in Texas test scores
      - Accusations of test-rigging
      - Houston failed to report half of their drop-outs

- **Teachers must be “highly qualified”**
  - BA
  - Pass a test

- **ADEQUATE YEARLY PROGRESS**
  - Determined by state
  - Raise achievement levels of African Americans, Latinos, low SES, and special Ed.
  - Everyone must be proficient by 2013-2014
**Adequate Yearly Progress**

- Uses standardized test scores + grad rates or unexcused absences
- If a single group within a school fails to reach proficiency, the entire school is seen as falling short
- Penalties are in place for failing to make AYP for schools receiving Title 1 funds

**Mandated testing** – progress will be measured annually in reading and math (grades 3 through 8), and at least once in high school.
- By end of 07/08 school year science will be tested 3 times (3-5, 6-9, 10-12)
- Aimed at increasing accountability in education, and closing the achievement gap

**AYP Not Met**

- 2 years means they are “in need of improvement”
  - Parents can send children to another public school in the district
  - Develop an improvement plan to be approved by the district
- 3 years
  - Add tutoring and after school programs
- 4 years
  - Corrective action (removing relevant staff, implementing new curriculum, decreasing management authority, using outside experts to advise, extend length of the school day, restructure internal organization of school)
- 5 years
  - Begin planning for restructuring
- 6 years
  - Replace all staff, state takeover

**Current Status of testing in Washington**

- State Superintendent Randy Dorn (elected 2009) dropped the WASL as Washington’s test.
- Instead, a shift was made to:
  - Grades 3-8 Measurements of Student Progress
  - High School Proficiency Exam
- This test can be conducted on-line, is shorter.

**Critiques**

- Encourages teaching to the test
- Each state produces their own standardized tests
  - No standardization across states
  - Arizona’s instrument to measure standards has been criticized for being too easy
- Cultural bias: Test scores reflect school performance, but also other inequalities in opportunities that exist in our society
  - Ten percent of white children live in poverty, while about 35 percent of Black and Latino children live in poverty.
  - Special Ed and ELL still have to be proficient in 2013
  - 1st Year ELL are given a “bye” on the reading test for one year after moving to the U.S.
EOC Assessment

- RCW 28A.655.061
- Math HSPE is started being replaced with End-of-Course (EOC) assessments for Algebra I and Geometry.
- Shifts the focus back onto classroom assessment (assumes we should already be testing whether students are learning what we teach them)

Current Status

- Identify 4 types of schools:
  - Reward (Top 10% of AYP for 3 years)
  - Priority (Lowest 5% overall)
  - Focus (Lowest 10% for subgroup)
  - Emerging (next 5 or 10% from Priority or Focus)
- Focus and Priority schools are assisted in taking steps to improve
- NO AYP penalties
- http://www.k12.wa.us/ESEA/PublicNotice.aspx

Smarter Balanced

- Currently under development
- Uses Computer Adaptive Technology
- Accessability
  - Braille
  - ASL videos
  - Pop-up Spanish glossary for math
- First year of full use will be 2014/15

NCLB Waiver

- In July 2012, Washington State received a waiver from the U.S. Department of Education for some Elementary and Secondary Education Act (ESEA) requirements
  - The sanctions for not meeting AYP are no longer in effect

Current Tests

- Joined the Smarter Balanced Assessment Consortium, to create a new assessment focused on the Common Core State Standards (vs. EALRs)
- http://www.smarterbalanced.org/about/members-states/
- Testing transitioning toward this test:
  - http://www.k12.wa.us/assessment/statetesting/

Technical Report

“The information in these reports (scaled score, performance levels, and standard score indicators) can be used with other assessment information to help with school, district, and state curriculum planning and classroom instructional decisions. High school students are required to pass the HSPE Reading and Writing and to participate in the Mathematics assessments as part of the graduation requirements.

While school and district scores may be useful in curriculum and instructional planning, it is important to exercise extreme caution when interpreting individual reports. The items included on MSP/HSPE tests are samples from a larger content domain. Scores from one test given on a single occasion should never be used solely to make important decisions about students’ placement, the type of instruction they receive, or retention in a given grade level in school. It is important to consider individual scores on MSP/HSPE tests with classroom-based and other local evidence of student learning (e.g., scores from district testing programs). When making decisions about individuals, multiple sources of information should be used. Multiple individuals who are familiar with the student’s progress and achievements—including parents, teachers, school counselors, school psychologists, special education teachers, and perhaps the students themselves—should be brought together to collaboratively make such decisions.”
The MSP/HSPE is currently administered by ETS.
They appear to have kept much of the WASL structure intact.
WASL was previously administered by Riverside, and later Pearson.

MSP/HSPE Construction

1) Selected content committees of 10 to 25 people (mostly teachers and curriculum experts) for each grade level.
2) Reviewed the EALRs: Essential Academic Learning Requirements
   - Content areas for grades K-10
3) Prepared test specifications (weight, importance, time)
   - Purpose of the Assessment
   - Strands (subtests)
   - Item Types (multiple choice, short answer, and extended response)
   - General Considerations of Testing Time and Style
   - Test Scoring
   - Distribution of Test Items by Item Type

4) Item writers (mostly teachers in an item writing workshop) create items. Items refined by content specialists (14 or more years experience in classroom teaching)
5) Sample items reviewed by content committees
   - Appropriate to EALRs?
   - Appropriate difficulty?
   - One correct response if MC, appropriate scoring guidelines for open response
6) 7) Reviewed by Bias & Fairness review committee for offensive or inappropriate stereotypes about gender, ethnicity, or culture
   - "Breaking Through" is a fictionalized memoir by Francisco Jiménez, an author and professor at Santa Clara University in California. The main character is a youth who works summers in the strawberry fields with his father for low wages — about $1 per hour. Their employer comes to visit their house and makes a reference to another worker being deported.
8) Item Pilot – Each test form contains 5 to 7 piloted items. Approximately 1200 responses per item
   - Analyze data with CCT and IRT – remove or revise and re-pilot bad items
   - Differential Item Functioning – does the item curve have different thetas across groups?
9) To create a new form of the test, items are selected from the pool based on the test blueprint.
10) Test is equated, using IRT, to the original baseline to maintain score interpretability
    - Math standards re-set in 2008
    - Others use 2004 standards

Old method of determining standards

Standard Setting Committee – 14 members, mostly math and LA teachers
Judges received:
   - items ranked from least to most difficult (though they did not know how many students answered each question correctly)
   - A scoring rubric
   - Judges thought about what made each item more difficult than the next.
   - Judges were instructed to flag “meets standard”, “exceeds standard”, and one between “near standard” (partially proficient) and low – a student at meets standard should answer 2/3 of preceding questions correctly;
   - Judges shared flags, discussed disputed items
   - Judges independently re-flagged, etc. - several iterations of this
New method of determining standards

- Panelists take test
- Panelist shown performance level descriptors and contrasting group study results
- Round 1 rating by panelists
- Panelists given % students answering each item correctly
- Round 2 rating by panelists
- Panelists presented with the percentage of students in each performance category?
- Round 3 rating by panelists

Technical report cites a standard setting document that is not yet available

Standards

- In all content areas, the standard (Level 3) reflects what a well taught, hard working student should know and be able to do.
- NOT MINIMAL COMPETENCY! Not what should be necessary for graduation from HS
- WEAP will advocate for and/or support an assessment of what a "typical" adult needs to manage a successful life in the 21st Century, then advocate for and/or support the adjustment of the assessment process at each testing level, and, by implication, 12th grade graduation requirements accordingly.

Level 4 - Advanced (previously Above Standard): This level represents superior performance, notably above that required for meeting the standard at grade 10.
  - 425 or greater

Level 3 - Proficient (previously Meets Standard): This level represents solid academic performance for grade 10. Students reaching this level have demonstrated proficiency over challenging content, including subject-matter knowledge, application of such knowledge to real world situations, and analytical skills appropriate for the content and grade level.
  - 400 or greater

Level 2 - Basic (previously Below Standard): This level denotes partial accomplishment of the knowledge and skills that are fundamental for meeting the standard at grade 10.
  - 375 or greater

Level 1 - Below Basic (previously Well Below Standard): This level denotes little or no demonstration of the prerequisite knowledge and skills that are fundamental for meeting the standard at grade 10.

MSP/HSPE

- Reliabilities
  - .81 to .91 for reading, math, and science, depending on grade level
  - Writing is .77
  - Fairly equivalent across genders and ethnicities (Table 6.1 & 6.2)
  - More variance in special program groups (particularly gifted)
  - Inter-rater reliabilities for writing: 75% to 89% "exact agreement"

- SEMs
  - Width of SEMs depends on score (Page 90). Wider at extreme ends.
  - Conditional reliability is the reliability of individuals receiving a given score.

- Predictive validity: No studies of success;
  - UW found that WASL scores predicted Fresh GPA the same as the SAT on ACT

- Content validity established through EALRs and Content Committees
  - an independent study suggests some tests items don't match the EALRS

- Construct validity:
  - WASL Correlations with ITBS suggest they are measuring similar constructs (no data on MSP/HSPE yet)
  - High correlations between Math and reading (math involves lots of reading)
  - Score differences across groups are as expected (gifted kids outperform others, ELLs score worse on reading and writing, etc.)

- Strand Scores:
  - Reliabilities vary widely, from .35 to .77 (not useful for anything at the individual level – Tables 6.24 to 6.30)
  - Strand inter-correlations generally fit what would be expected (Tables 5.2 to 5.11)
  - Factor analysis suggests two factors: reading/writing and math. Science cross-loads on both
Personality Tests

- **Informal Assessment** – Observations, interviews
- **Structured Personality Inventories** – Neo, MMPI
- **Projective Techniques**: Rorschach, TAT, Rotter

Structured Personality Inventories

- **Content-related procedures**: Define what is being measured and develop items that reflect the content of interest. This makes for a very face-valid measure, that is easy to fake, or easy for a person to respond to similar items in similar ways.
- **Personality theory** – use theory as a basis for the instrument, determine if the measure meets the tenets of the theory; NEO
- **Factor Analysis** – groupings of items are determined by FA, 16 PF (Cattell; lexical approach, hasn’t been replicated)
- **Empirical criterion-keying**: Select items based on an external criterion, MMPI

Common Personality Measures

- **Minnesota Multiphasic Personality Inventory (MMPI)**
  - A very well-researched measure of psychopathology
  - Used for clinical assessment
- **Millon Clinical Multiphasic Inventory (MCMI)**
  - Focuses on personality disorders
- **NEO Personality Inventory – Revised**
  - Based on the Five factor Model of personality
  - Well-regarded measure of normal personality

- **Myers-Briggs Type Indicator (MBTI)**
  - Based on Jungian theory, divides people into “types”
  - Widely used in I/O settings for team-building exercises
  - Horrible psychometric properties – between 39 to 76% of people have different types when re-tested
- **Rorschach Inkblot test**
  - Projective test, originally had no administration or interpretation instructions
  - Now uses the Exner System

Other Projectives

- **Thematic Apperception Test**
- **Draw-a-Person**
- **House-Tree-Person**

Please, if you are using a test, investigate the research on the validity of the test for the specific purpose!

MMPI

- Early personality inventory: Woodworth personal data sheet (t/f/ symptom list, no theoretical or empirical basis used for item inclusion); Used in WWI to determine emotionally unsuitable recruits
- Hathaway & McKinley (in early 1930s) objected to the “rational” approach taken to measuring personality. Developed using **empirical criterion-keying** (items are keyed to an external criterion).
- MMPI – Minnesota Multiphasic Personality Inventory
MMPI

1. Assembled 1000 items from psychiatric textbooks, personality inventories, and clinical experience
2. Reduced item pool (duplicates and "insignificant" items) to 504.
3. Classified items under 25 headings (social attitudes, morale, etc.)
4. Normative group – 724 individuals who were friends or relatives of patients at the University Hospital in Minneapolis (fairly representative of Minnesota in the 30s – WHITE)
5. 4 additional normative groups (265 precollege students, 265 skilled workers, 254 non-psychiatric patients, 221 psychiatric patients)
6. Select a criterion group (50 cases of pure, uncomplicated hypochondriasis)
7. Items were included if the proportion of respondents between criterion and normative groups were statistic different
   1. Items eliminated if criterion group’s response was less than 10% (deviant responses excluded because they describe few respondents)
   2. Items eliminated if related to marital status or socioeconomic status
8. The other normative groups were used to modify the scales

MMPI-2

■ Restandardization of the MMPI: 567 T/F items
■ Some items deleted due to content (mostly GI and religiously based), some items re-worded
■ Added 154 provisional items
■ 704 items
■ Normative sample: 2,600 people in different states (chart for demographics) (double-checked with a census-matched subsample)
■ The vast amount of research conducted on the MMPI has changed the way it is used.
   ■ We know what is related to high and low scores on many clinical scales
   ■ A large number of supplemental scales have also been developed

MMPI-2 Clinical Scales

■ Scale 1 — Hypochondriasis
   ■ Neurotic concern over bodily functioning.
   ■ Criterion group: 50 uncomplicated hypochondriacal patients
■ Scale 2 — Depression
   ■ Poor morale, lack of hope in the future, and a general dissatisfaction with one’s own life situation.
   ■ High scores are clinical depression whilst lower scores are more general unhappiness with life.
   ■ Criterion group: 50 uncomplicated cases of bipolar in a depressive phase

■ Scale 3 — Hysteria
   ■ Hysterical reaction to stressful situations. Often have "normal" facade and then go to pieces when faced with a 'trigger' level of stress.
   ■ People who tend to score higher include brighter, better educated and from higher social classes. Women score higher too.
   ■ Criterion group: 50 patients diagnosed with hysteria or identifiable histrionic personality components.
   ■ Scales 1, 2, and 3 are sometimes called the “Neurotic Triad”

■ Scale 4 — Psychopathic Deviation
   ■ Measures social deviation, lack of acceptance of authority, amorality.
   ■ Adolescents tend to score higher.
   ■ Criterion group: Young people (17 to 22), diagnosed with psychopathic personalities, avoid social and amoral types, referred to court due to delinquent activities (not major criminal activities)
■ Scale 5 — Masculinity-Femininity
   ■ Originally developed to identify homosexuals, but did not do so accurately.
   ■ Measures how strongly an individual identifies with the traditional (pre-1960's) masculine or feminine role.
   ■ Men tend to get higher scores. It is also related to intelligence, education, and socioeconomic status.
   ■ Criterion group: 13 male homosexual inverts – lots of attempts to come up with criterion groups.
MMPI-2 Clinical Scales

- Scale 6 — Paranoia
  - Paranoid symptoms such as ideas of reference, feelings of persecution, grandiose self-concepts, suspiciousness, excessive sensitivity, and rigid opinions and attitudes.
  - Criterion group: Exact number and content not known; paranoid state, paranoid condition, paranoid schizophrenia

- Scale 7 — Psychasthenia
  - Originally characterized by excessive doubts, compulsions, obsessions, and unreasonable fears, it now indicates conditions such as Obsessive Compulsive Disorder (OCD).
  - It also shows abnormal fears, self-criticism, difficulties in concentration, and guilt feelings.
  - Criterion group: 20 patients diagnosed with psychasthenia

- Scale 8 — Schizophrenia
  - Assesses a wide variety of content areas, including bizarre thought processes and peculiar perceptions, social alienation, poor familial relationships, difficulties in concentration and impulse control, lack of deep interests, disturbing questions of self-worth and self-identity, and sexual difficulties.
  - Criterion group: 2 groups of 50 schizophrenics
  - Scales 6, 7, and 8 are often called the “Psychotic triad”

- Scale 9 — Hypomania
  - Tests for elevated mood, accelerated speech and motor activity, irritability, flight of ideas, and brief periods of depression.
  - Criterion group: 24 manic patients with mild or moderate mania.

- Scale 0 — Social Introversion
  - Tests for a person’s tendency to withdraw from social contacts and responsibilities.
  - Criterion group: 50 female students performing above the 65th percentile on a measure of introversion/extraversion, 50 female students scoring below the 35th percentile.

MMPI-2 Validity Scales

- The L scale
  - Originally called the "Lie" scale, this was an attempt to assess naive or unsophisticated attempts by people to present themselves in an overly favorable light.
  - These items were rationally derived rather than criterion keyed.

- The F scale
  - “InFrequent”, this is a deviant, or rare response scale.
  - These are items which are rarely endorsed by normal people.
    - If less than 10 percent of the normals endorse the item, but you do, your F count goes up.
    - "All laws should be eliminated."
  - Other F scales have been developed: F(b), F(p)
### MMPI-2 Validity Scales

- **The K scale**
  - An attempt to assess clinically defensive responses.
  - Constructed by comparing the responses of
    - a group of people who were known to be clinically deviant but who produced normal MMPI profiles to
    - a group of normal people who produced normal MMPI profiles
  - The K scale is used to alter scores on other MMPI scales.
    - High K people give scores on other scales which are too low.
    - K-corrected and uncorrected scores are available when the test results are interpreted.

### MMPI-2

- In addition to the clinical and validity scales, there are many additional content and supplemental scales, as well as subscales for the major clinical scales.
- Research on the MMPI has given us a vast amount of information about the characteristics of individuals who receive different scores on the MMPI.
- **Concerns:**
  - Does not seem to over-pathologize ethnic groups (the original MMPI did).
  - Most concerns focus on the use of the test.
    - Electronic test scoring services provide automated score interpretations (clinicians need to take into account contextual info).
    - Test scores are sometimes used for personnel screening, in the absence of evidence linking scores to job performance.

### Key Concepts from MMPI

- **Empirical criterion keying** for item selection can (but doesn’t always) result in a less face-valid measure than would otherwise be possible.
- **Validity indices** can be used to guide interpretation of other, more substantive scales.
- The true utility of a measure lies in its ability to connect individual scores to research findings.

### Career Counseling

- **Strong Interest Inventory**: 317 items – rate whether they are like, indifferent to, or dislike a number of occupations, school subjects, people, etc.
- People who are satisfied with a given profession share a number of interests and characteristics.
- **General Occupational Themes**: Realistic, social, investigative, enterprising, artistic, conventional.
- **Occupational Scales**: Rates whether interests are similar to those doing the job.
- **Personal Style Scale**: Work style, learning environment, leadership style, risk taking/adventure.

### Strong Interest Inventory

- 211 occupational scales (separate scales for men and women for occupations dominated by gender – plumbers, child-care).
- Items were generated to describe occupational interests.
- Like the MMPI, the SII used Empirical Scale Construction:
  - Data were collected from samples of successful workers for each occupation.
  - Items that discriminated one occupational group from the remaining samples were considered on that occupational scale.
  - Empirical scale construction does not insure predictive validity!
    - Just because two groups tend to differ from one another does not mean that what they differ on can be used to accurately predict...
Strong Interest Inventory

- Test-retest reliabilities range from .80-.90 with intervals from 2 weeks to 3 years
- Lower test-retest reliabilities with HS students than college students or working adults
- Occupational interests are more malleable the younger the respondent
- Cronbach’s alpha is in low .90s

Strong Interest Inventory

- Little psychometric data has been generated by independent researchers
- SII correlates with choice of major
  - 80% of college students had selected a major that was at least moderately related to their interests as measured by the SII – BUT IT ISN'T NECESSARILY THE HIGHEST CORRELATION
- SII correlates with choice of profession
  - Musicians score higher on “performing arts”
  - Chemists score high on “science”
  - Elementary school teachers score higher on “teaching and education”

Strong Interest Inventory

- Bizot & Goldman (1993) found that interest inventories administered in HS predicted neither job performance nor job satisfaction 8 years later
- Other research has found correlations between SII and job satisfaction

IQ Tests

- 1930s - David Wechsler – Studies under Spearman and Pearson, was a tester in WWI
  - Wechsler was interested in intelligence as a series of inter-related factors. Specifically, he was interested in differences between Army alpha (verbal) and army beta (performance).
- 1939 - Wechsler-Bellevue Intelligence Scale
- 1946 Form II
- 1949 Wechsler Intelligence Scale for Children (downward extension of form II)

WISC-IV

- Verbal Comprehension Index – Verbal knowledge and understanding gained through formal and informal education
- Perceptual Reasoning Index – Ability to interpret and organize visual information and solve problems
- Working Memory Index – Ability to temporarily retain information in memory and manipulate it; attention, concentration, and mental control
- Processing Speed Index – Ability to quickly and correctly scan and sequence visual information; concentration and grapho-motor skills
WISC-IV

- The new version of the WISC-IV:
  - Verbal Comprehension Index (VCI): Similarities, Vocabulary, and Comprehension subtests;
    - verbal concept formation, verbal reasoning, knowledge acquired from one's environment;
    - low scores may be due to impaired language skills, closed head injury, raised in a deprived/chaotic environment
  - Perceptual Reasoning Index (PRI): Block Design, Picture Concepts, and Matrix Reasoning subtests;
    - perceptual and fluid reasoning, spatial processing, visual-motor integration;
    - low score may be low due/to associated skills (VMI) nonverbal learning disability, Turner's syndrome

- Working Memory Index (WMI): Digit Span and Letter-Number Sequencing subtests;
  - temporarily retain info in memory, perform some operation or manipulation with it, and produce a result, attention, concentration, mental control;
  - Results may be due to reading disorder, language disorders, risk factor for learning difficulties

- Processing Speed Index (PSI): Coding and Symbol Search subtests.
  - Ability to quickly and correctly scan, sequence, or discriminate simple visual information; ST visual memory; attention, visual-motor coordination;
  - associated with general cognitive ability – faster processing conserves mental resources; ADHD, learning disability, TBI

WISC-IV

- Supplemental Scales: Up to one supplemental can be substituted for a core test for each of the indices:
  - VCI (Information and Word Reasoning),
  - PRI (Picture Completion),
  - WMI (Arithmetic), and
  - PSI (Cancellation).

WISC-IV

- Content Validity: Content validity was based on extensive literature reviews, and input from panels, consultants, and various psychologists.
- Item Bias
  - National tryout phase (n = 1,270 oversampled of 252 African American and 186 Hispanic children)
  - Items were reviewed by "content and bias experts"
  - Also looked at differential item functioning (DIF) –
    - Statistical tests for determining bias (different IRT curves, % of groups answering an item correct)
    - Not enough info on comparison groups and criteria for removing items

WISC-IV

- Range of ability tested
  - Assessed with item difficulty curves;
  - Very solid, cover a range of ability levels;
  - Scaled scores greater than 2/less than 2 SDs from the mean can be obtained at all age levels

WISC-IV

- Standardization Sample
  - 2,200 examinees divided into age groups (200 in each) (only 1100 for Arithmetic)
  - Representative of the March 2000 U.S. Bureau of the Census data in terms of age, gender, race ethnicity, parent education level, and four geographical regions.
  - Large areas within regions exclude examinees, without clarification (8 sites in Idaho, only 1 in Nevada)
**WISC-IV**

- **Reliabilities:** Generally quite high – page 34
  - Subtest reliability ranges from .79 to .9 across ages
  - Index reliabilities from .88 to .94 across ages
  - Full Scale IQ .97 for all years but 6 & 7 (%)
- **Factor Structure:** The four factor model is replicated best when optional subtests are not included. Strong evidence from Psych Corp.
  - There is also good evidence for a 5-factor model (more in line with contemporary research):
    - Crystallized ability
    - Visual processing
    - Fluid reasoning
    - Short-term memory
    - Processing speed

- **Convergent/Divergent Validity**
  - FSIQ correlated .89 to WISC-III, WPPSI-III, and WAIS-III; .66 to WASH (remember these correlations are attenuated by reliability)
  - Based on 244 examinees with somewhat different demographic characteristics than the WISC-IV standardization sample, the WISC-IV and WISC-III were correlated .89 (with an average FSIQ decrease of 2.3 points – Flynn Effect).

<table>
<thead>
<tr>
<th>Trait Method</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Verbal VCI</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. Other</td>
<td>.83</td>
<td>-</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. Percep PRI</td>
<td>.62</td>
<td>.61</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>4. Other</td>
<td>.61</td>
<td>.76</td>
<td>-</td>
<td></td>
</tr>
</tbody>
</table>

  - The correlations between index scores and subtests match what would be expected (page 51)

- **Diagnostic Utility**
  - Useful for diagnosing Mental Retardation and Giftedness
  - Using for assessing traumatic brain injury if a premorbid assessment is available (not so in these studies)
  - Not useful by itself for diagnosing learning disabilities (too much within-group variation); You need MANY SOURCES OF INFORMATION
  - Not useful for assessing ADHD (though it is an important part of an ADHD assessment)

- **Group studies and norms available for (p152 – 174):**
  - Reading, Written Expression, and mathematics disorders
  - ADHD
  - Expressive Language D/O; Mixed Receptive-expressive
  - TBI
  - Autism
  - Aspergers
  - These are based on small sample sizes, no standardization of diagnosis

- **Administration:**
  - Administered 1-to-1 by a doctoral level psychologist (or qualified technician)
  - Manual lists 65-80 minutes for the core subtests ;
    - Gifted average 104
    - Normal 94
    - MR average 73
  - Item administration for a scale is discontinued when a specified number of item scores of 0 occur.
    - These discontinue rules were specified by determining at what point fewer than 2% of children in normative samples gave no additional correct responses after a given sequence of errors.

- **Interpretation**
  - Best method is not that described by Psych Corp. Rather by Flanagan and Kaufman (2004):
  1. Analysis of the WISC-IV indexes to determine the best way to summarize a child’s overall intellectual ability.
  2. Normative (compared to sample) and Personal (compared to self) Strengths and Weaknesses among the indexes are identified. Relative to both same-age peers from the normal population (normative approach) and the child’s own overall ability level (individual or self-evaluation approach).
  3. Optional interpretive steps involving new WISC-IV composites to uncover additional information about the child’s cognitive capabilities and generate potentially meaningful hypotheses about areas of integrity or dysfunction.
WAIS-III

- **WAIS-III**
- 9 core subtests, 2 optional for dichotomy, 2 optional for 4-factor:
- Can be scored/interpreted 2 ways:
  - VIQ PIQ dichotomy, or
  - 4 factors:
    - Verbal Orientation Index (VOI)
    - Perceptual Organization Index (POI)
    - Working-Memory Index (WMI)
    - Processing Speed Index (PSI)
- FSIQ ranges from 45 to 155
- Now paired with the Wechsler memory scale

WAIS-III

- **Standardization:**
- **Sample:**
  - 2450 people, stratified into 13 age groups
  - Matches 1995 census data with respect to gender, socioeconomic status, race and ethnicity, educational attainment, and geographical residence.
- **Purpose** is to assess psychoeducational disability, neuropsychiatric and organic dysfunction, and giftedness.

WAIS-III

- **Studies show profiles of people with:**
  - Alzheimer's, Huntington's, and Parkinson's diseases;
  - temporal lobe epilepsy;
  - traumatic brain injury; and
  - Korsakoff's syndrome,
  - schizophrenia, and
  - people with psychoeducational and developmental disorders (mental retardation, attention-deficit-hyperactivity, learning disorders, and deaf and hearing impaired)

Other IQ Tests

- **Stanford-Binet; Largely fallen out of favor**
- **Kaufman Brief Intelligence Test (K-BIT): Based on modern intelligence theory**
- **Kaufman’s tests are widely used**
- **Raven’s Progressive Matrices**
- Uses only matrix reasoning
- **Uniform Nonverbal Intelligence Test (UNIT)**
- Developed to be “language free”

IQ Hodge-podge

- **IQ is partly hereditary (40 to 80%) – roughly 50%:**
- **Correlations between:**
  - The same person tested twice: 87%
  - Identical twins reared together: 86%
  - Identical twins reared apart: 76%
  - Fraternal twins reared together: 55%
  - Biological siblings: 47%
  - Parents and children living together: 40%
  - Parents and children living apart: 31%
  - Adopted children living together: 0%
  - Unrelated people living apart: 0%

Mental Retardation

- **IQ score ranges (from DSM-IV):**
  - mild mental retardation: IQ 50–55 to 70; children require mild support; formally called "Educable Mentally Retarded".
  - moderate retardation: IQ 35–40 to 50–55; children require moderate supervision and assistance; formally called "Trainable Mentally Retarded".
  - severe mental retardation: IQ 20–25 to 35–40; can be taught basic life skills and simple tasks with supervision.
  - profound mental retardation: IQ below 20–25; usually caused by a neurological condition; require constant care.
- **Males have higher rates (no extra protective X chromosome)**
Brain size: Modern studies using MRI imaging have shown that brain size correlates with IQ ($r = 0.35$) among adults (McDaniel, 2005).

Flynn Effect: IQ scores worldwide appear to be slowly rising at a rate of around three IQ points per decade (Flynn, 1999). Attempted explanations have included:
- Improved nutrition, a trend towards smaller families, better education, greater environmental complexity, and business (Mingroni, 2004).
- May be slowing down in some 1st world countries

Gender: men and women have the same average IQ. Between-group differences are smaller than within group differences:
- Women perform better on tests of memory and verbal proficiency for example, while
- Men perform better on tests of mathematical and spatial ability.
- Although gender-related differences in average IQ are insignificant, male scores display a higher variance: there are more men than women with both very high and very low IQs.

Flynn Effect: IQ scores worldwide appear to be slowly rising at a rate of around three IQ points per decade (Flynn, 1999). Attempted explanations have included:
- Improved nutrition, a trend towards smaller families, better education, greater environmental complexity, and business (Mingroni, 2004).
- May be slowing down in some 1st world countries

Gender: men and women have the same average IQ. Between-group differences are smaller than within group differences:
- Women perform better on tests of memory and verbal proficiency for example, while
- Men perform better on tests of mathematical and spatial ability.
- Although gender-related differences in average IQ are insignificant, male scores display a higher variance: there are more men than women with both very high and very low IQs.

Race: East Asian > White > Sub-Saharan African / Hispanic;
- Not completely removed by taking into account SES;
- Lots of controversy!!

Health: Persons with a higher IQ have generally lower adult morbidity and mortality.
- This may be because they better avoid injury and take better care of their own health,
- or alternatively may be due to a slight increased propensity for material wealth (see above).
- Post-Traumatic Stress Disorder, severe depression, and schizophrenia are less prevalent in higher IQ bands.

Achievement tests
- Measures of academic achievement, by subject
- Typically norm-based (NOT criterion referenced)
- Wide Range Achievement Test (WRAT-R)
- Woodcock Johnson PsychoEducational Battery

Wide Range Achievement Test (WRAT-R)
- Measures of academic achievement, by subject
- Typically norm-based (NOT criterion referenced)
- Wide Range Achievement Test (WRAT-R)
- Woodcock Johnson PsychoEducational Battery

Woodcock Johnson
- 10 subtests that can be scored into –
  - Reading, Mathematics, Written language
  - Skills, Application, Fluency
- Standardized on 4,732 people from 3 to 80, census matched gender, race, occupational status, geographic region and urban/nonurban
- Achievement reliabilities in the .80s and .90s
- Validity coefficients range from .4 to .8

Race: East Asian > White > Sub-Saharan African / Hispanic;
- Not completely removed by taking into account SES;
- Lots of controversy!!

Health: Persons with a higher IQ have generally lower adult morbidity and mortality.
- This may be because they better avoid injury and take better care of their own health,
- or alternatively may be due to a slight increased propensity for material wealth (see above).
- Post-Traumatic Stress Disorder, severe depression, and schizophrenia are less prevalent in higher IQ bands.

Achievement tests
- Measures of academic achievement, by subject
- Typically norm-based (NOT criterion referenced)
- Wide Range Achievement Test (WRAT-R)
- Woodcock Johnson PsychoEducational Battery

Wide Range Achievement Test (WRAT-R)
- Measures of academic achievement, by subject
- Typically norm-based (NOT criterion referenced)
- Wide Range Achievement Test (WRAT-R)
- Woodcock Johnson PsychoEducational Battery

Woodcock Johnson
- 10 subtests that can be scored into –
  - Reading, Mathematics, Written language
  - Skills, Application, Fluency
- Standardized on 4,732 people from 3 to 80, census matched gender, race, occupational status, geographic region and urban/nonurban
- Achievement reliabilities in the .80s and .90s
- Validity coefficients range from .4 to .8

Gender: men and women have the same average IQ. Between-group differences are smaller than within group differences:
- Women perform better on tests of memory and verbal proficiency for example, while
- Men perform better on tests of mathematical and spatial ability.
- Although gender-related differences in average IQ are insignificant, male scores display a higher variance: there are more men than women with both very high and very low IQs
Age/Grade Equivalent Scores

- Used most often with achievement tests
- Age equivalent is sometimes called “mental age”
- Common interpretation of grade equivalent scores:
  - 7.0, The student is performing at a 7th grade level
  - 3.5, The student is performing as well as someone halfway through with the 3rd grade

Problems with Grade Equivalents

- Grade Equivalent scores are ordinal, not continuous
  - Grade equivalents assume that growth is equally distributed over the year
  - Within a distribution, intervals do not represent equal units (ordinal data)
  - Many equivalent scores are obtained by extrapolation, not actually obtained from children's scores
  - Encourages disparate groups to be compared to one another
  - A 2nd grader with a grade equivalent of 4.1 is not the same as a 4th grader (they only got the same number of items correct)
  - Grade equivalents exaggerate small differences (a 2-grade level difference may be well within the norms of a given grade level)
  - Generally, avoid grade equivalents – They don’t tell you much about an individual’s performances, and are easily misinterpreted.

Test Bias

- A test produces biased scores when people with the same ability perform differently on a test due to their affiliation with a particular group.
  - Group differences do not mean bias
  - Tests can tend to produce biased scores, but scores are biased

- Content Bias – Content of the test is more familiar to one group than another, and that familiarity influences the scores
  - Words or expressions that are associated with a particular group
  - Lack of clarity in instructions
  - Assessed with Differential Item Functioning (DIF)

- Bias in Internal Structure – When psychometric properties are inconsistent across groups
  - Factor structure
  - Reliability

- Selection Bias – When a test varies in predictive validity across groups
Other Sources of Bias

- Language and reading level
- Test-wiseness (how experienced of a test-taker is the individual?)
- Motivation and anxiety
- Stereotype threat – A loss of self-regulatory ability when a stereotype is made salient.
- Examiner bias

IDEA

- Public Law 94-142 (Education of All handicapped Children Act) passed in 1975, now codified as IDEA (Individuals with Disabilities Education Act)
- Individualized Educational Plans (IEPs) fall under IDEA
  - 504s fall under civil rights laws

Learning disability:

- Broadly – Learning difficulties associated with any type of factor (including MR, brain injury, sensory difficulties, or emotional disturbance)
- Narrowly – The failure to learn a scholastic skill on the part of a child with adequate intelligence, maturity, cultural background, and educational experience

Specific learning disability (PL 94-142) – a disorder in one or more of the basic psychological processes involved in understanding or in using language, spoken or written, which may manifest itself in an imperfect ability to listen, think, speak, read, write, spell, or do mathematical calculations.

- INCLUDES: brain injury, dyslexia, developmental aphasia;
- DOES NOT INCLUDE: problems that are primarily the result of visual, hearing, or motor handicaps, MR, emotional disturbance, or of environmental, cultural, or economic advantage.

National Joint Committee for Learning Disabilities:

- I.D is a generic term that refers to a heterogeneous group of disorders manifested by difficulties in acquisition and use of listening, speaking, reading, writing, reasoning, or mathematical abilities;
- Presumed to be due to central nervous system dysfunction; not a direct result (though can be concurrent with) emotional disturbances, sensory impairment, etc.

4 – stage Information processing Model

Receptive (input)

- Sensory storage – short-term sensory store preserves representation of image for a brief time
- Perceptual encoding – initial representation is encoded into a more permanent representation; held in STM for transfer to LTM
- Central Processing – Encoded information manipulated, compared to others in STM, associations with LTM
- Response selection mechanisms - Make decision of how to respond

Expressive (output)
Children with LD have difficulty with active information processing:
- Don’t use as many mnemonic aids (rehearsal, labeling, chunking, selective attention)
- Don’t attend to and remember central information as well as their peers
- Reading D/O – difficulty encoding phonological information in LTM
- Overall – difficulty with selective attention and analyzing the task in a way that results in the best performance strategy

Assessing LD

- PL 94-142
  - The child does not achieve commensurate with his or her age and ability levels in one or more of the areas (when provided with appropriate learning experiences)
  - Severe discrepancy between achievement and intellectual ability in one or more of the following areas: oral expression, listening comprehension, etc.
  - Not a specific learning disability if ability-achievement discrepancy is the result of visual, hearing, or motor handicap, MR, emotional disturbance, or environ, etc. disadvantage.

Assessing LD

- Indicator
  - Most frequent indicator is ability-achievement discrepancy
  - The Law does not indicate how large the discrepancy should be
  - Not appropriate: Low achievement (regardless of intelligence) may result from factors not having to do with learning
  - Not appropriate: Discrepancy between verbal and performance sections of an IQ test (does not take into account achievement in school tasks)

Defining Discrepancies

- Look to see at which grade level the student is performing; specify a number of grade levels that are defined as discrepant
  - not good – 2 years below is much worse for a 4th grader than an 8th grader
  - A number for each grade level (graduated deviation criterion)
- Expectancy formula – Expected grade equivalent: \( \frac{(2MA + CA)}{3} - 5 \); Relies heavily on the concept of “mental age” – little theoretical support
- Difference between standard scores – This implies a near-perfect correlation between ability and achievement; make sure each are based on the same standard score
- Regression – best method – takes into account reliability of instruments and expected correlation between them

Washington Discrepancy Tables

- Uses FSIQ, if FSIQ is not an appropriate estimate of ability, deference is given to the judgment of a qualified professional
- Consider minimum intellectual ability – must be above what would be considered for MR
  - if FSIQ not an accurate measure, deference is given to the judgment of a qualified professional
- Be careful with minority groups or early grades, as measures may not be reliable
- Cannot be applied below grade level 1
- Tests not listed should have reliability of .85 or greater, nationally normed, validity coefficients available

Response to Intervention

- A new paradigm, gaining increasing popularity among school administrators.
  - Children with academic difficulties are identified.
  - The teacher modifies the curriculum, instruction, etc.
  - If the child responds to the intervention, it is held in place. If not, a new intervention is tried.
  - This escalates until the problem is solved
RTI

Pros:
- All children with academic difficulties can benefit
- Interventions put in place more quickly

Cons:
- Does not identify LD, particularly in high IQ groups
- Does not allow for gifted education (this depends on the district – more schools are allowing this).
- Special education requires specialized training
- Puts onus of implementing individualized educational plans on general classroom teachers