# Week 6 - Estimating validity (part 1)

Thomas J. Faulkenberry, Ph.D.

A test is **reliable** if observed score variability is due to *true score variability*

A test is **valid** if it measures what it purports to measure

- Three types of validity
  - Content validity - subjective rating of items by content experts
  - Criterion-related validity (this week)
  - Construct validity (next week)

## Criterion-related validity

- test scores are related to some *criterion*
  - Ex: a test for screening job applicants relates to the criterion of *work effectiveness*
  - Ex: a school admissions test relates to the criterion of GPA (or % of students who complete program)

- in simple cases, reliability measured with a *correlation coefficient*
  - $\rho_{XY}$, where $X =$ test score and $Y =$ criterion score
  - "validity coefficient"

What role does reliability play?

- a test cannot correlate more highly with any other score than it correlates with *its own true score*
  - $\rho_{XY} \leq \rho_{XT} = \sqrt{\rho_{XX'}}$
- if a test does not have perfect reliability (i.e., $\rho_{XX'} < 1$), then the validity coefficient is **attenuated** (reduced)
  - that is, $\rho_{XY} < \rho_{T_X T_Y}$

Spearman (1904) proposed the following idea to *correct* for this attenuation.

First, note that, by definition, $\rho_{XY} = \dfrac{c_{XY}}{\sigma_X \sigma_Y}$.

It is easy to prove that $c_{XY} = c_{T_X T_Y}$. Thus, we have $\rho_{XY} = \dfrac{c_{T_X T_Y}}{\sigma_X \sigma_Y}$.

So we can compute the correlation between true scores as:

$$
\begin{aligned}
\rho_{T_X T_Y} = \frac{c_{T_X T_Y}}{\sigma_{T_X}\sigma_{T_Y}} &= \frac{\rho_{XY}\sigma_X\sigma_Y}{\sigma_{T_X}\sigma_{T_Y}} \\
&= \frac{\rho_{XY}}{\left(\frac{\sigma_{T_X}}{\sigma_X}\right)\left(\frac{\sigma_{T_Y}}{\sigma_Y}\right)} \\
&= \frac{\rho_{XY}}{\sqrt{\rho_{XX'}}\sqrt{\rho_{YY'}}}
\end{aligned}
$$

Example: Suppose Test A relates to Criterion C with correlation 0.4 and Test B relates to Criterion C with correlation 0.3. Further suppose that reliability of A = 0.81, reliability of B = 0.25, and reliability of C = 0.64. Which test (A or B) is the more valid predictor of Criterion C?

Example: Suppose Test A relates to Criterion C with correlation 0.4 and Test B relates to Criterion C with correlation 0.3. Further suppose that reliability of A = 0.81, reliability of B = 0.25, and reliability of C = 0.64. Which test (A or B) is the more valid predictor of Criterion C?

Let's compute the "corrected" validity coefficients:

- Test A: $\rho_{T_A T_C} = \dfrac{\rho_{AC}}{\sqrt{\rho_{AA'}}\sqrt{\rho_{CC'}}} = \dfrac{0.4}{\sqrt{0.81}\sqrt{0.64}} = 0.56$

- Test B: $\rho_{T_B T_C} = \dfrac{\rho_{BC}}{\sqrt{\rho_{BB'}}\sqrt{\rho_{CC'}}} = \dfrac{0.3}{\sqrt{0.25}\sqrt{0.64}} = 0.75$

Example: Suppose Test A relates to Criterion C with correlation 0.4 and Test B relates to Criterion C with correlation 0.3. Further suppose that reliability of A = 0.81, reliability of B = 0.25, and reliability of C = 0.64. Which test (A or B) would be better for predicting criterion C?

Let's compute the "corrected" validity coefficients:

- Test A: $\rho_{T_A T_C} = \dfrac{\rho_{AC}}{\sqrt{\rho_{AA'}}\sqrt{\rho_{CC'}}} = \dfrac{0.4}{\sqrt{0.81}\sqrt{0.64}} = 0.56$
- Test B: $\rho_{T_B T_C} = \dfrac{\rho_{BC}}{\sqrt{\rho_{BB'}}\sqrt{\rho_{CC'}}} = \dfrac{0.3}{\sqrt{0.25}\sqrt{0.64}} = 0.75$

Thus, Test B has the largest *potential* validity coefficient

- if we make the test more reliable by adding more items

Note that the formula works for correcting attenuation in single variables too:

For example, if you wanted to estimate the true correlation between a predictor *observed* score $X$ and a criterion *true* score $Y$, you would just "partial out" the reliability of the criterion – that is,

$$\rho_{XT_Y} = \frac{\rho_{XY}}{\sqrt{\rho_{YY'}}}$$

Often, criterion-related tests are used for screening people into **dichotomous categories**

- hire / do not hire
- successful / not successful

Often, criterion-related tests are used for screening people into **dichotomous categories**

- hire / do not hire
- successful / not successful

Basic procedure:

1. set a "cut score" $Z$ on the criterion variable $Y$
2. if $Y \geq Z$, place into one category ("successful")
3. if $Y < Z$, place into other category ("unsuccessful")

Example: Suppose 100 job applicants are screened with a test.

| | | Prediction Successful | Unsuccessful | Total |
|---|---|---|---|---|
| Actual outcome | Successful | 30 | 20 | *50* |
| | Unsuccessful | 10 | 40 | *50* |
| | Total | *40* | *60* | *100* |

Example: Suppose 100 job applicants are screened with a test.

| | | Prediction | | |
|---|---|---|---|---|
| | | Successful | Unsuccessful | Total |
| Actual outcome | Successful | 30 | 20 | *50* |
| | Unsuccessful | 10 | 40 | *50* |
| | Total | *40* | *60* | *100* |

Validity for such a test is related to the "hit rate"

- i.e., proportion of applicants correctly screened (prediction=outcome)

- for this example, hit rate $= \dfrac{30+40}{100} = 0.70$

Example: Suppose 100 job applicants are screened with a test.

| | | Prediction | | |
|---|---|---|---|---|
| | | Successful | Unsuccessful | Total |
| Actual outcome | Successful | 30 | 20 | *50* |
| | Unsuccessful | 10 | 40 | *50* |
| | Total | *40* | *60* | *100* |

Validity for such a test is related to the "hit rate"

- note that this hit rate is "good", because it works better than just hiring everybody (success rate of 50%)

- here, 50% is the "base rate" - the best hit rate available from all possible procedures

Example: Suppose 100 people are screened for some psychopathology:

| | | Prediction | | |
|---|---|---|---|---|
| | | Psychopathology | No Psychopathology | Total |
| Actual | Psychopathology | 7 | 3 | *10* |
| | No psychopathology | 14 | 76 | *90* |
| | Total | *21* | *79* | *100* |

Example: Suppose 100 people are screened for some psychopathology:

| | | Prediction | | |
|---|---|---|---|---|
| | | Psychopathology | No Psychopathology | Total |
| Actual | Psychopathology | 7 | 3 | *10* |
| | No psychopathology | 14 | 76 | *90* |
| | Total | *21* | *79* | *100* |

- hit rate $= \dfrac{7+76}{100} = 0.83$
- but base rate = 0.90 (we could just categorize everyone as "no psychopathology")
- better off giving no test at all!

How to set cut scores?

Consider the following score distribution. Which cut score would give us the largest hit rate?

| Score | Successful | Unsuccessful |
|---|---|---|
| 20 | 3 | 0 |
| 19 | 5 | 0 |
| 18 | 12 | 2 |
| 17 | 8 | 1 |
| 16 | 10 | 2 |
| 15 | 4 | 5 |
| 14 | 1 | 8 |
| 13 | 1 | 10 |
| 12 | 2 | 7 |
| 11 | 1 | 5 |
| 10 | 1 | 4 |
| 9 | 0 | 3 |
| 8 | 0 | 2 |

Let's just try a few:

**Cut score = 15**

| | | Prediction | | |
|---|---|---|---|---|
| | | Successful | Unsuccessful | Total |
| Actual outcome | Successful | | | *48* |
| | Unsuccessful | | | *49* |
| | Total | | | *97* |

Idea:

- we are predicting that scores **at or above 15** will be successful.
- count how many of those *successful*
- count how many are *unsuccessful*

Let's just try a few:

**Cut score = 15**

| | | Prediction | | |
|---|---|---|---|---|
| | | Successful | Unsuccessful | Total |
| Actual outcome | Successful | 42 | | *48* |
| | Unsuccessful | | | *49* |
| | Total | | | *97* |

Idea:

- we are predicting that scores **at or above 15** will be successful.
- count how many of those *successful*: 42
- count how many are *unsuccessful*

Let's just try a few:

**Cut score = 15**

| | | Prediction Successful | Unsuccessful | Total |
|---|---|---|---|---|
| Actual outcome | Successful | 42 | | *48* |
| | Unsuccessful | 10 | | *49* |
| | Total | | | *97* |

Idea:

- we are predicting that scores **at or above 15** will be successful.
- count how many of those *successful*: 42
- count how many are *unsuccessful*: 10

Let's just try a few:

**Cut score = 15**

| | | Prediction Successful | Unsuccessful | Total |
|---|---|---|---|---|
| Actual outcome | Successful | 42 | 6 | *48* |
| | Unsuccessful | 10 | 39 | *49* |
| | Total | 52 | 45 | *97* |

Idea:

- calculate the remaining cells
- Hit rate $= \dfrac{42+39}{97} = 0.835$

* Let's just try a few:

**Cut score = 16**

| | | Prediction Successful | Unsuccessful | Total |
|---|---|---|---|---|
| Actual outcome | Successful | 38 | 10 | *48* |
| | Unsuccessful | 5 | 44 | *49* |
| | Total | 43 | 54 | *97* |

- Hit rate $= \dfrac{38 + 44}{97} = 0.845$

Let's just try a few:

**Cut score = 17**

| | | Prediction Successful | Unsuccessful | Total |
|---|---|---|---|---|
| Actual outcome | Successful | 28 | 20 | *48* |
| | Unsuccessful | 3 | 46 | *49* |
| | Total | 31 | 66 | *97* |

- Hit rate $= \dfrac{28+46}{97} = 0.763$
- So, setting cut score to **16** gives best hit rate (0.845)