

## Riverside Evalu8 2023

Multi-Language Test Development:  
A Deep Dive into the Norming Process

Beth Varner, M.Ed., NCSP, LEP #3925

1

## Agenda

- Norming
- Calibration
- Multi-Language Test Development
- Implications for Practice
- Q&A


3

## W Scale


- Special transformation of the Rasch ability scale
  - Mathematical framework used to measure latent traits such as attitude or ability
  - Shows the probability of an individual getting a correct response on a test item
    - Subject ability
    - Item difficulty
- W scale for each test – centered on a value of 500, which approximates the average performance of 10 year, 0 months (age norms) or 5<sup>th</sup> grade, 0 month (grade norms) individuals

5

## W Scale



- Equal-interval scale
- Difficulty levels of items
- Ability level that represents success on a test




6

## Using the W Scale: Norming

- Norm-referenced assessments allow us to compare the performance of a student to a representative sample of the US populations
- WJ IV Norming Study
  - 7,416 participants (preschool through adult)
  - Native English speakers

7

## Using the W Scale: Norming



When we administer the test to a large representative group of people, we obtain information about the average ability of people at different ages. We can then use that information to create norms.

8

### Using the W Scale: Norming

For example, on this test the average ability of examinees 8 years, 0 months is located at 478 on the w scale. The standard deviation is 10 points. (68% of the examinees age 8-0 have scores between 468 and 488.)

9

### Using the W Scale: Norming

On the same test, the average ability of examinees 10 years, 0 months is located at 500 on the W scale. The standard deviation is 16 points. (68% of the examinees age 10-0 have scores between 484 and 516.)

10

### Using the W Scale: Norming

Norms – the *set of distributions* that describe the performance of the examinees at each age or grade in the norming sample. As a whole, this set of distributions illustrate the *rate of growth for the ability* in the population. The underlying scale is the w scale. All of the examinee distributions (norms) use the w scale as the unit of measurement.

11

### Calibration

- Identifies items that are a poor fit
- Sorts items by difficulty level
- Assigns a W difficulty to each item
  - a numerical value on the W scale that reflects the difficulty of the item
- Generates a W ability for each raw score
  - a numerical value on the W scale that reflects the ability level of the examinee

14

### Calibration

Items serve as a means of helping us estimate an examinee's location on the w scale.

15

### Norming

Measuring the *people* and charting growth of skill/ability/trait

### Calibration

Measuring difficulty of *items* and plotting them on the w scale by difficulty level

16

### Co-Norming

- Assessments based on the same norm study
- Allows for direct comparisons with high degrees of confidence
  - WISC & WJ IV ACH vs WJ COG & WJ ACH
- Multi-language co-norming
  - Only possible when we can reasonably assume the same rate of growth for abilities in both languages
  - Example: WJ IV, Bateria IV, WMLS III are co-normed because they are all built on the WJ IV Norming Study

21

### Multi-Language Test Development

#### Translation

- Contain the same items in both languages
- Can assume same difficulty level in both languages
- Measure constructs that should not be susceptible to differences in difficulty due to the language of administration (ex. Fluid Reasoning, Quantitative Reasoning, Visual Processing, some Cognitive Processing Speed tasks)

#### Adaptation

- Tests measure same underlying trait or construct, but different items are selected
- Cannot assume same difficulty level in both languages
- Measure constructs with strong language components (ex. Comprehension-Knowledge, Reading and Writing, Long-Term Storage and Retrieval, Auditory Processing)

22

### Reasons to Adapt

- Differences in item difficulty
  - Example: WJ IV ACH Test 1: Letter-Word Identification (Form A) Item #62 is "silhouette" which has a w difficulty of ~510
  - Bateria IV APROV test Identificación de letras y palabras Item #50 is "silueta" which has a w difficulty of ~487
  - Spanish equivalent of the English word is 23 w points easier
- Construct coverage
  - Example: WJ IV ACH Test 3: Spelling and Bateria IV Prueba 3: Ortografía – it is important that the items in each form reflects the orthography of the respective language
- Cultural relevance
  - For Spanish-speaking, some WJ IV items may be culturally irrelevant or inappropriate

23

### Spanish Calibration Study

- Required for new tests or tests that contain new items
- Purposes of calibration study:
  - Establishes the difficulty level of each new Spanish item
  - Maps these items on the w scale
- Ex. For Bateria IV, 6 tests were in the calibration study
  - 2 parts of COG Phonological Processing: Word Access & Substitution
  - COG Story Recall
  - COG Nonword Repetition
  - ACH Oral Reading
  - ACH Reading Recall

25

### Spanish Calibration and Equating

- Choose items from the existing English pool that have reasonable Spanish counterparts to serve as *anchor items*.
- Translate the anchor items into Spanish.
- Develop additional Spanish items (at this point, the new item difficulties are unknown)
- Administer translated Spanish anchor items and new Spanish items to a large pool of examinees.
- Place the new Spanish items onto the scale of the English items.

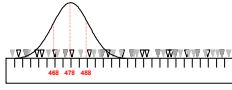
28

### Spanish Calibration and Equating

Now we have a w scale that contains items from both English and Spanish. We can locate an examinee's ability on this scale in either language, depending on which form of the test we administer.

29

### Spanish Calibration and Equating



The norms have not changed. For example, we still expect the average English-speaking student age 8-0 to score 478 on the English form. We also expect the average Spanish-speaking student age 8-0 to score 478 on the Spanish form.

Because the Spanish items have been placed onto the same scale as the English items, we can directly compare performance in the two languages for bilingual individuals.

30

### Implications for Practice

- Test selection
  - Informed selection balancing priorities of co-norming and representative samples
- Evaluation reports and eligibility meetings
  - Accurately describing the norming vs calibration samples, as needed

32

### Contact

**Beth Varner, M.Ed., NCSP**  
 School Psychologist  
 Licensed Educational Psychologist (CA – LEP #3925)

bethmvarner@gmail.com

bethvarner.com  
 squareholes.net

34