Considerations and Guidelines for Choosing and Using Screeners

YAACOV PETSCHER

FLORIDA STATE UNIVERSITY

FCRR
FLORIDA CENTER for READING RESEARCH

Main Ideas to Cover

- Why Screening?

- Building Blocks of Screener Evaluation

- Tools to Support Choosing and Using Screeners

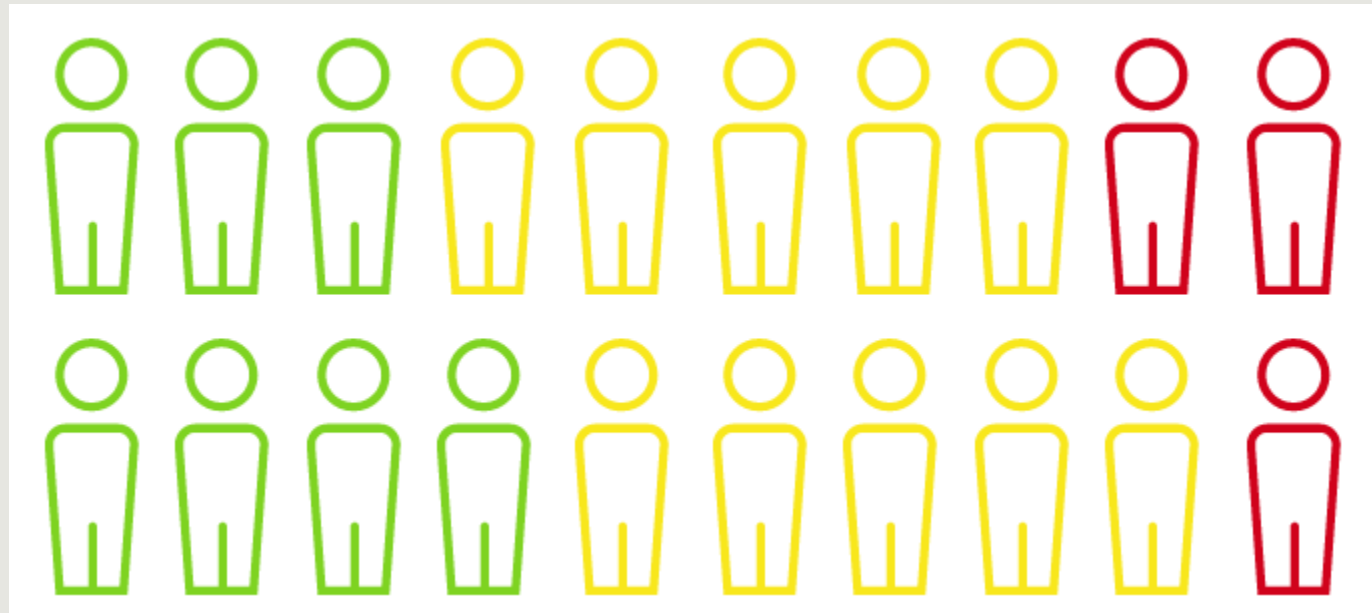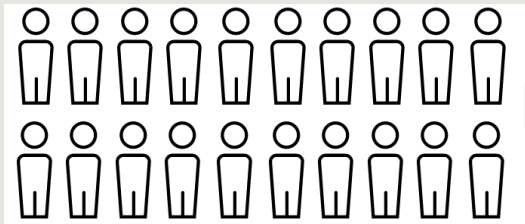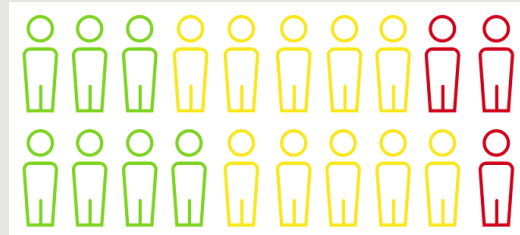# Why Screening?
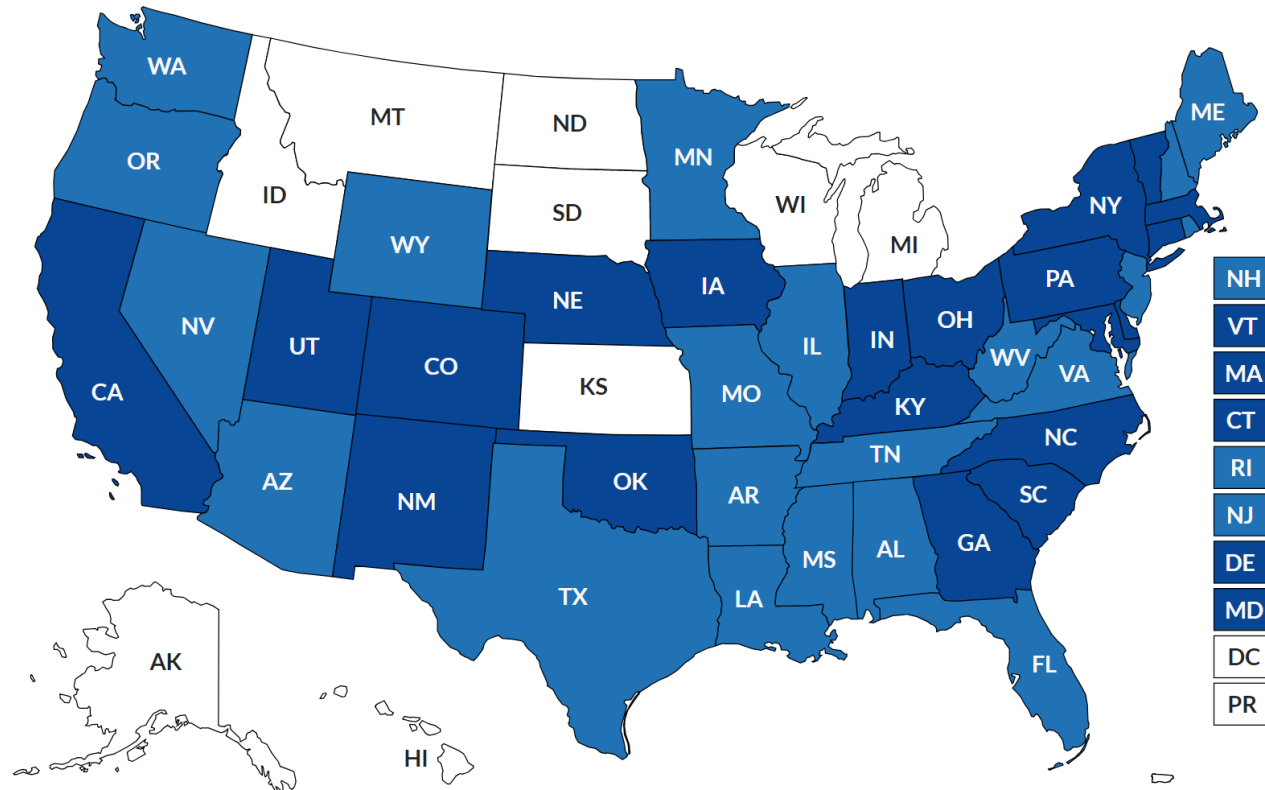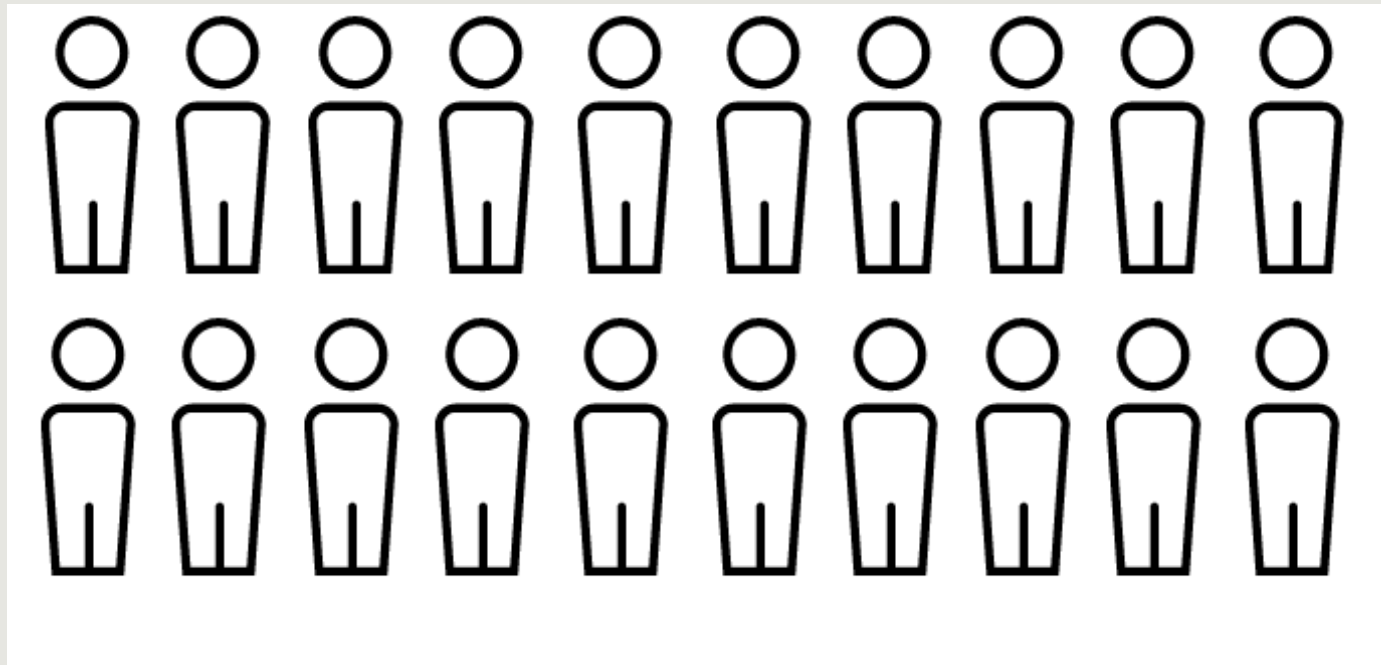
Why Screening?
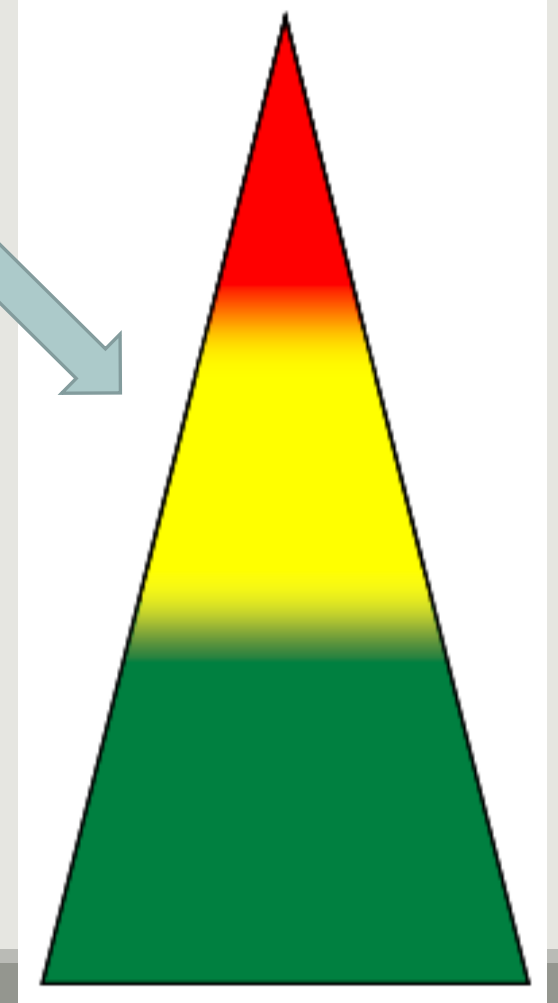
Why Screening?

"Rule out" rule
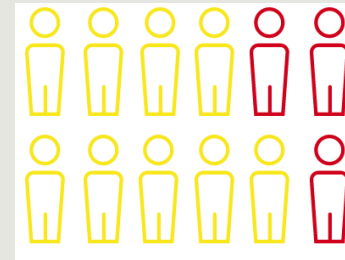
What kinds of screening routes?
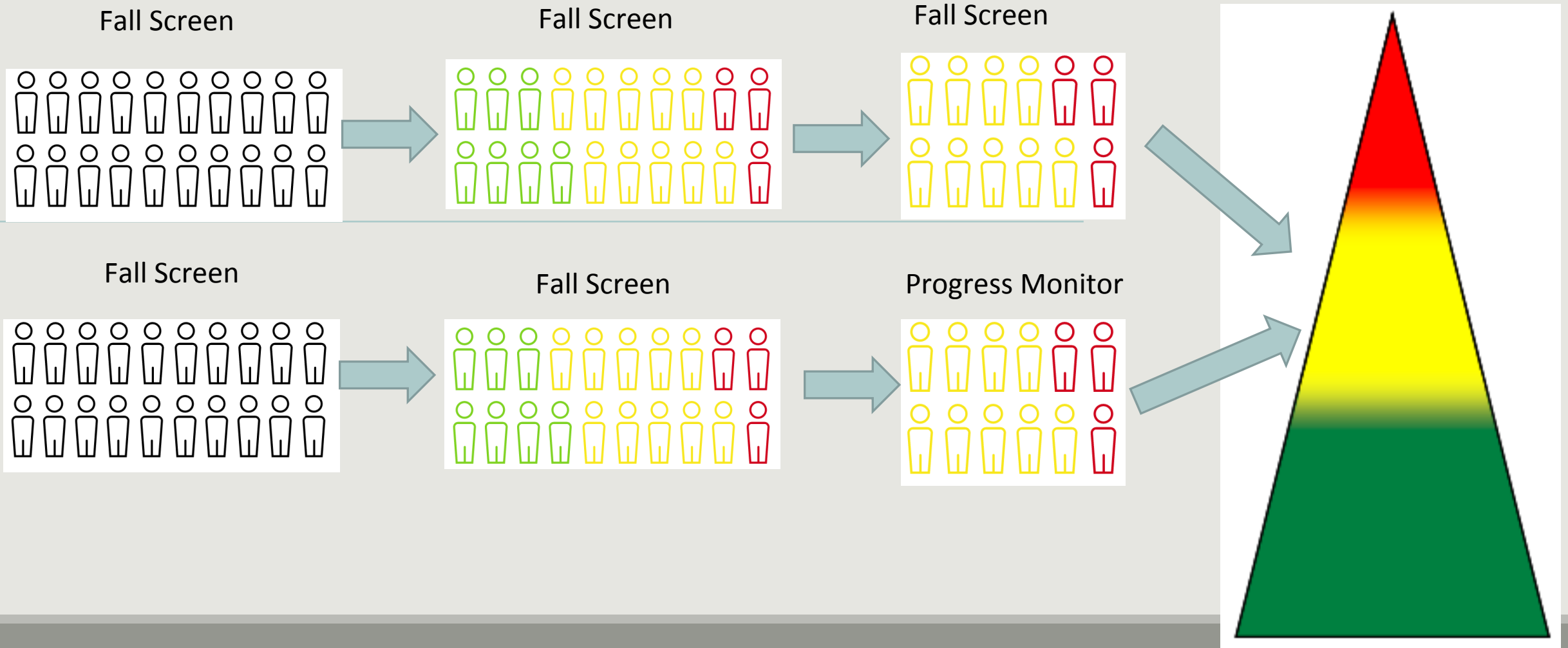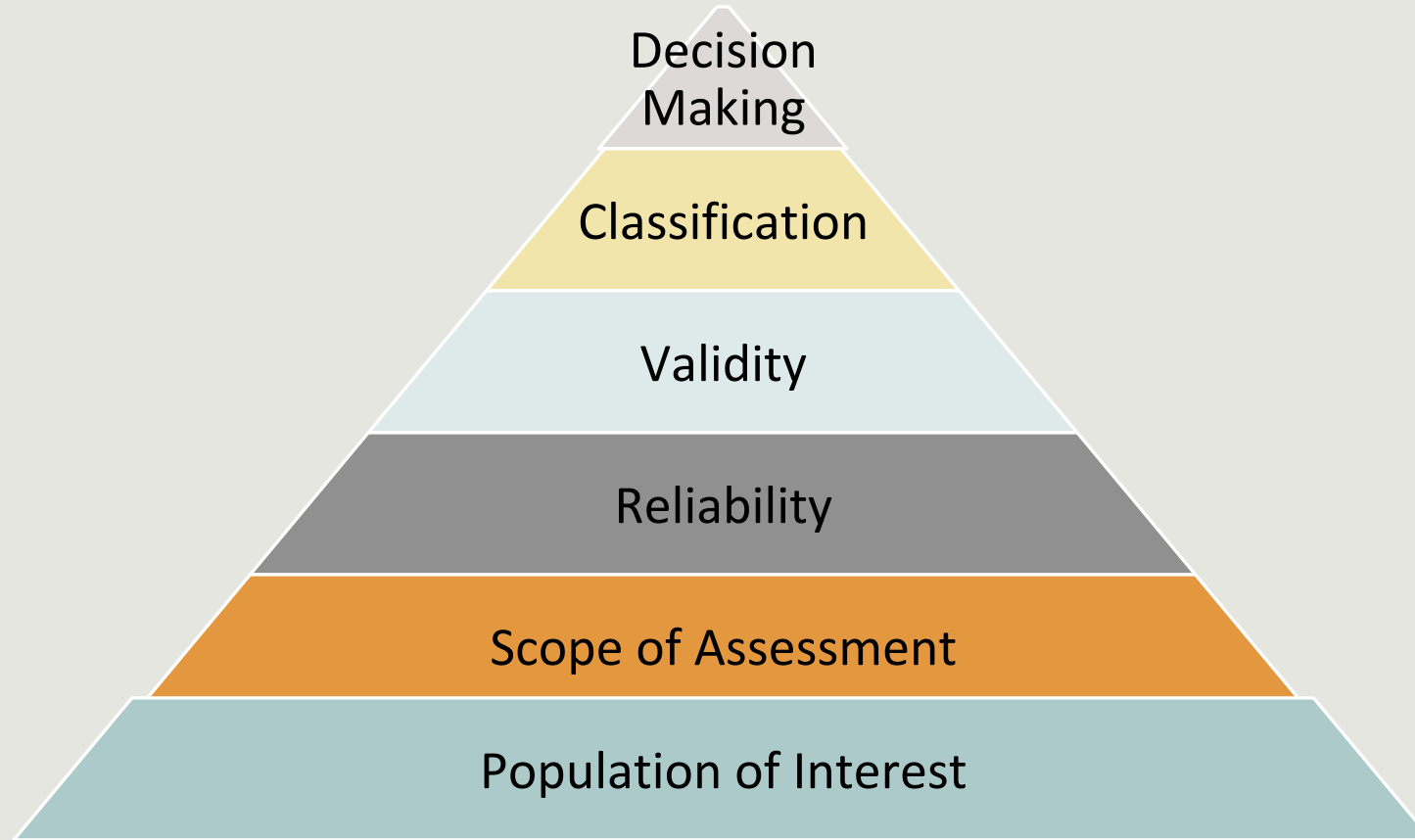

Fall Screen

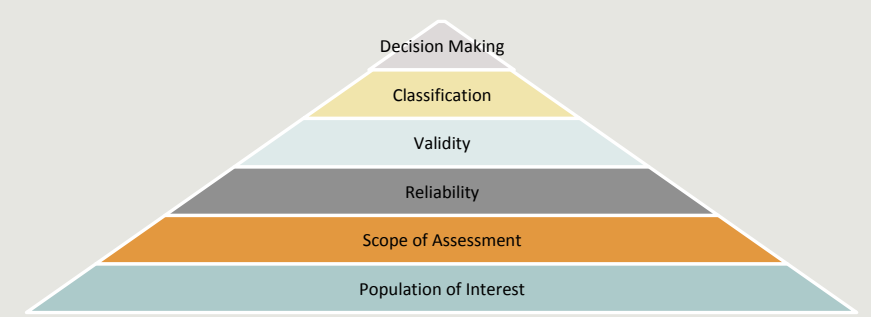Fall Screen

Fall Screen

What kinds of screening routes?

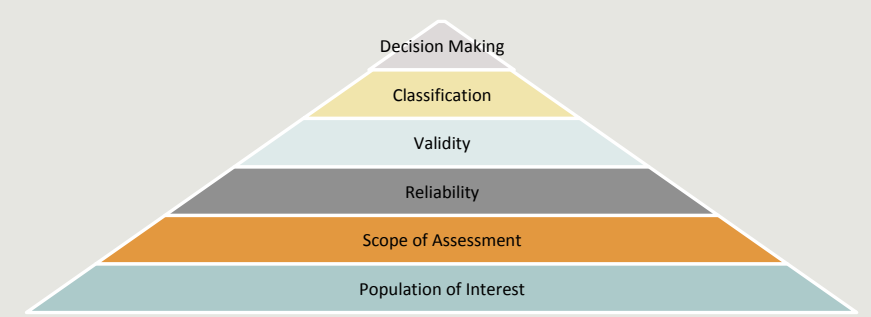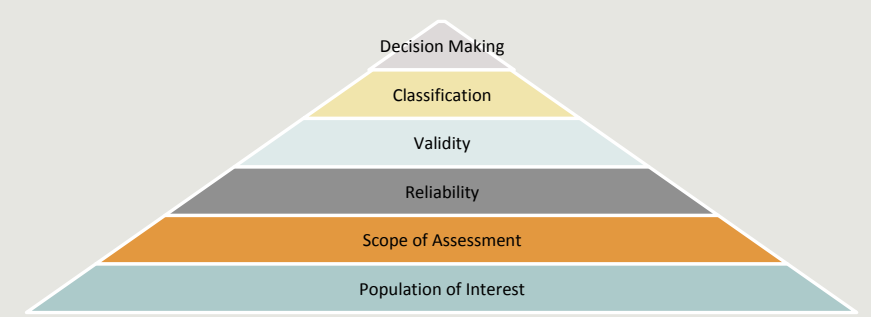# Problem Space – Building Blocks

Population of Interest

Children with dyslexia

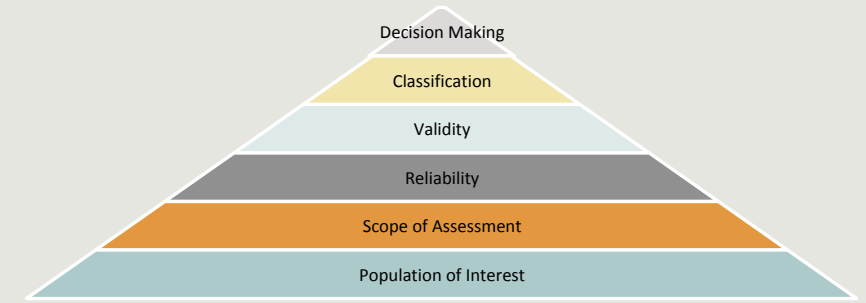# Population of Interest

Children with dyslexia

Population of Interest

Children with dyslexia

Population of Interest

IDA

"Dyslexia is a specific learning disability that is neurobiological in origin. It is characterized by difficulties with accurate and/or fluent word recognition and by poor spelling and decoding abilities. These difficulties typically result from a deficit in the phonological component of language that is often unexpected in relation to other cognitive abilities and the provision of effective classroom instruction. Secondary consequences may include problems in reading comprehension and reduced reading experience that can impede growth of vocabulary and background knowledge."

Mayo Clinic

"Dyslexia is a learning disorder that involves difficulty reading due to problems identifying speech sounds and learning how they relate to letters and words (decoding). Also called reading disability, dyslexia affects areas of the brain that process language. People with dyslexia have normal intelligence and usually have normal vision. Most children with dyslexia can succeed in school with tutoring or a specialized education program. Emotional support also plays an important role."

Scope of the Assessment – Screener/Outcome

Norm-Referenced Test

Criterion-Reference Test

Scope of the Assessment – Screener/Outcome
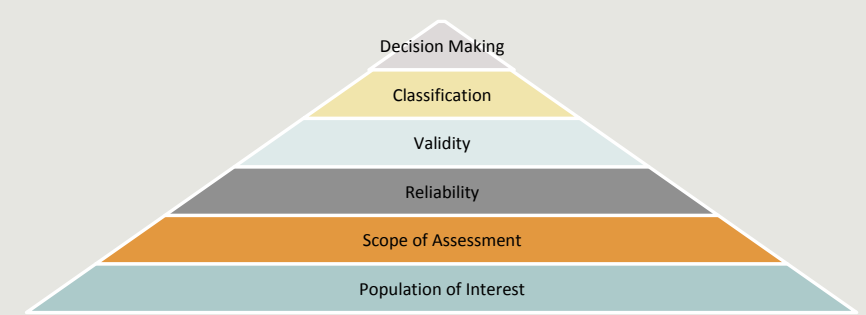
Table 4. Consistency of Comprehension Deficit Diagnosis (percentage of children diagnosed by both tests) for Full Sample and for Younger and Older Age Groups.

Test Pair	Full Sample Overlap Low 100	Younger Overlap Low 100 (20%)	Older Overlap Low 100 (20%)	Younger Overlap Low 50 (10%)	Older Overlap Low 50 (10%)
QRI-WJPC	50	54	46	52	38
PIAT-WJPC	56	67	55	64	50
PIAT-QRI	45	51	47	52	34
PIAT-GORT	35	43	51	36	32
QRI-GORT	36	41	47	40	36
WJPC-GORT	39	49	44	40	40

Note: GORT= Gray Oral Reading Test; PIAT = Peabody Individual Achievement Test; QRI = Qualitative Reading Inventory; WJPC= Woodcock-Johnson Passage Comprehension.

Scope of the Assessment – Speed/Power

Stats Warning!!

Math

$\sqrt{\ }$ $r_{xx}\, r_{yy} = r_{xy}$

$\sqrt{\ }$ $.75(.85) = .798$

$\sqrt{\ }$ $.70(.70) = .70$

Scope of the Assessment – Speed/Power

Reliability Differences Demonstrated

Reliability

Decision Making

Classification

Validity

Reliability

Scope of Assessment

Population of Interest
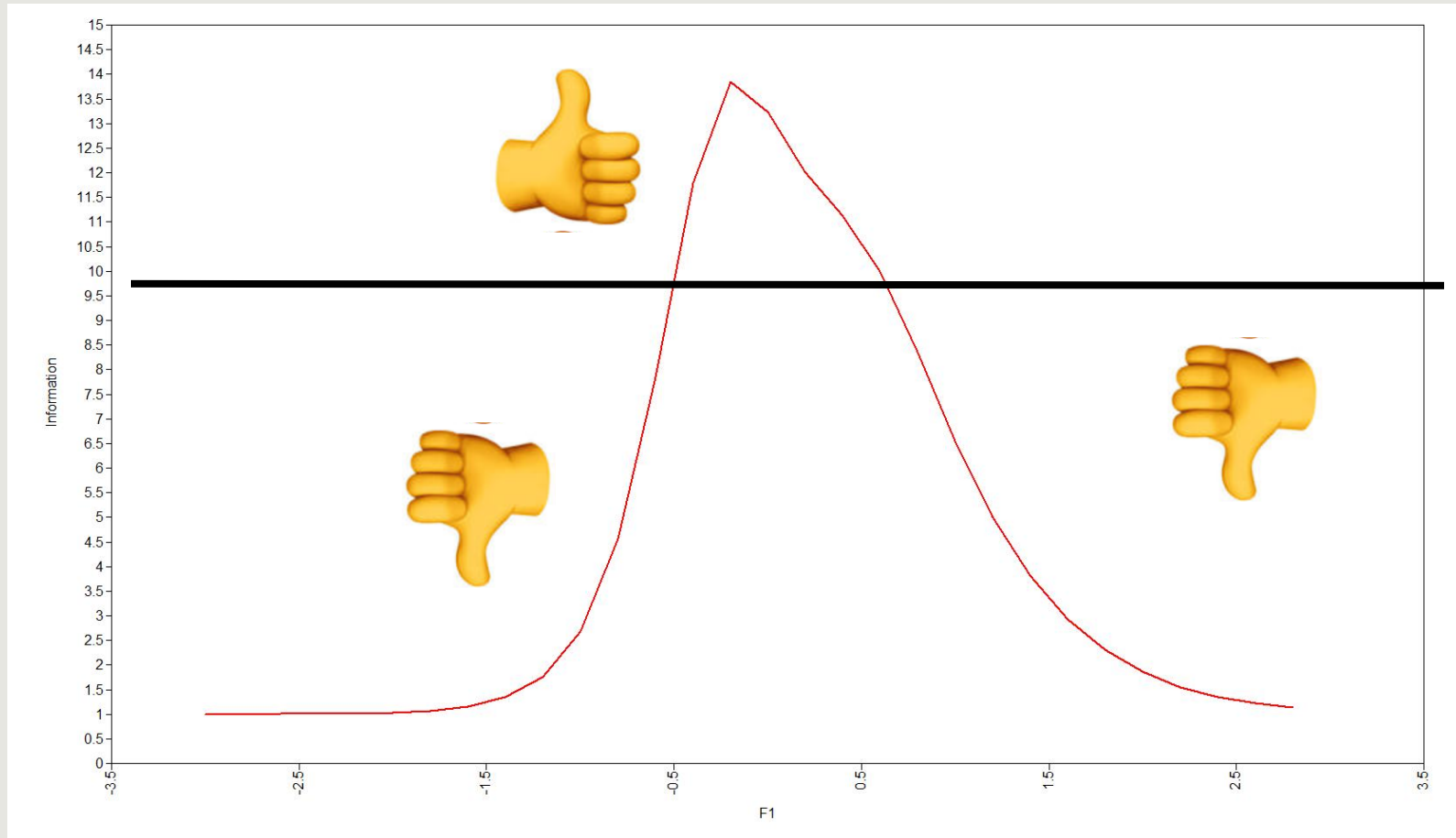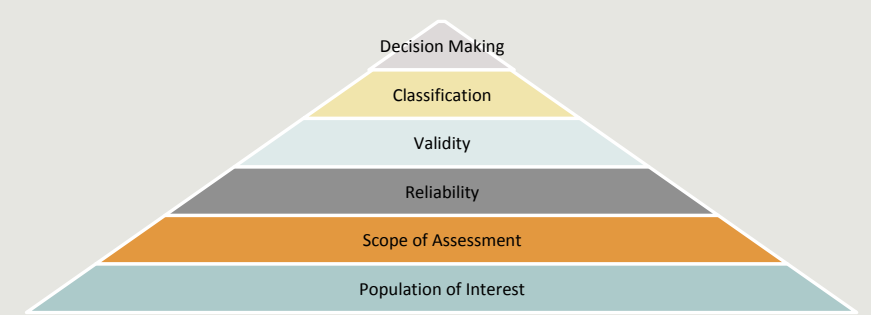
What is reliability?
- Stability, accuracy, consistency of scores
- Many types
  - Internal consistency
  - Alternate form
  - Retest
  - Split-half
- Not interchangeable!

Validity
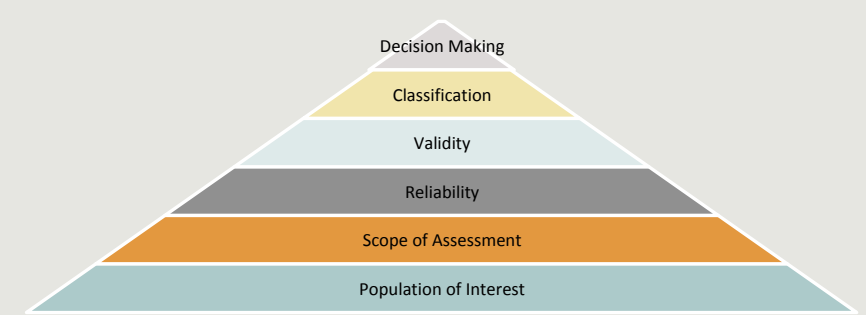
What isn't validity?
- Aetiological
- Conclusion
- Concurrent
- Consequential
- Construct
- Content
- Convergent
- Criterion
- Discriminant
- Ecological
- External
- Face
- Factor
- Generalizability
- Hypothesis

Even more
- In Situ
- Internal
- Nomological
- Predictive
- Structural
- Substantive
- Translational
- Treatment
- Washback

Confluence of Reliability & Validity

Classification

<table>
<tr><td></td><td></td><td colspan="2">Outcome</td></tr>
<tr><td></td><td></td><td><10th%ile</td><td>>=10th%ile</td></tr>
<tr><td rowspan="2">Screener</td><td>At-Risk</td><td>A</td><td>B</td></tr>
<tr><td>Not At-Risk</td><td>C</td><td>D</td></tr>
</table>
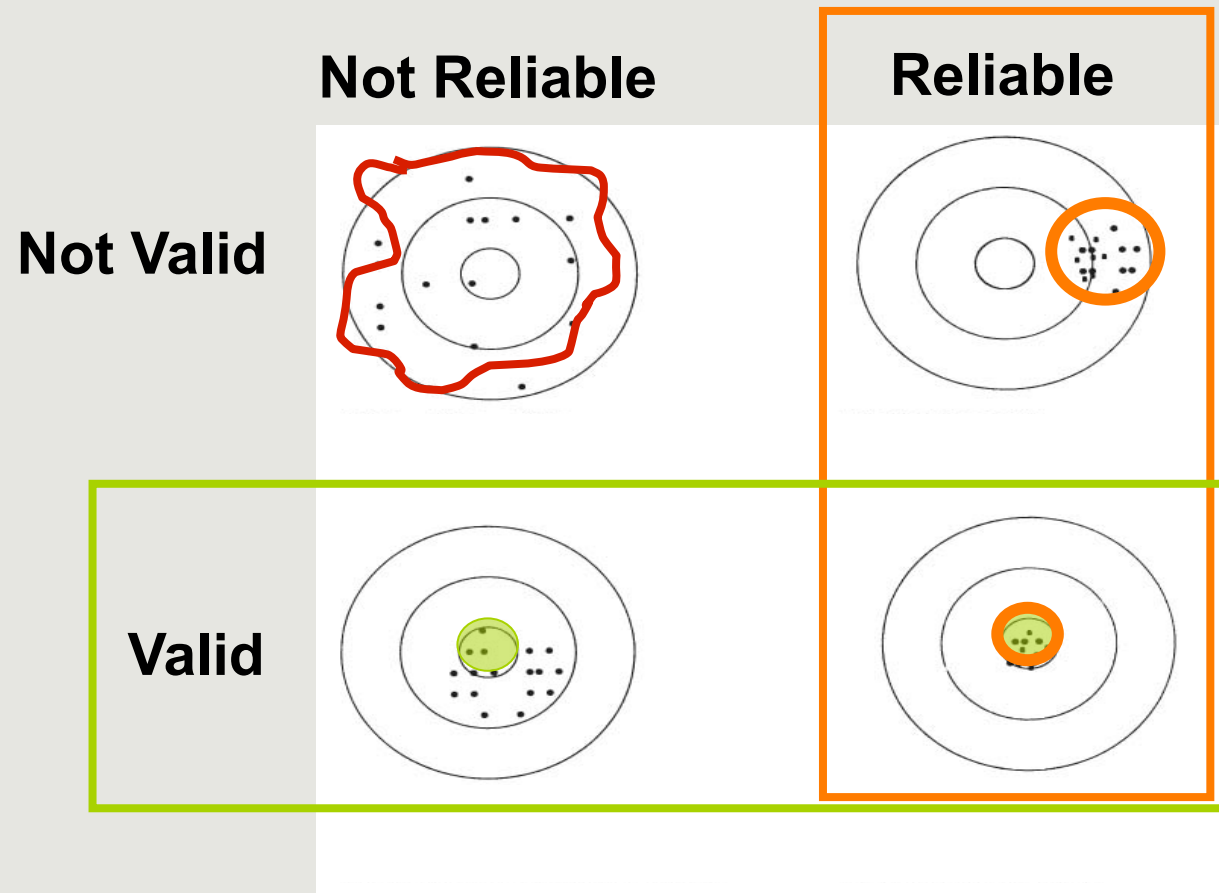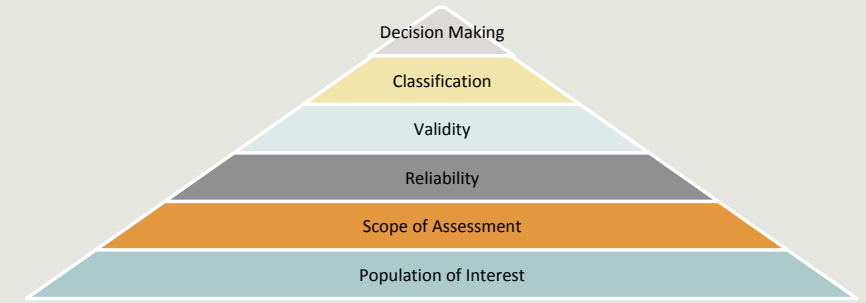
Classification
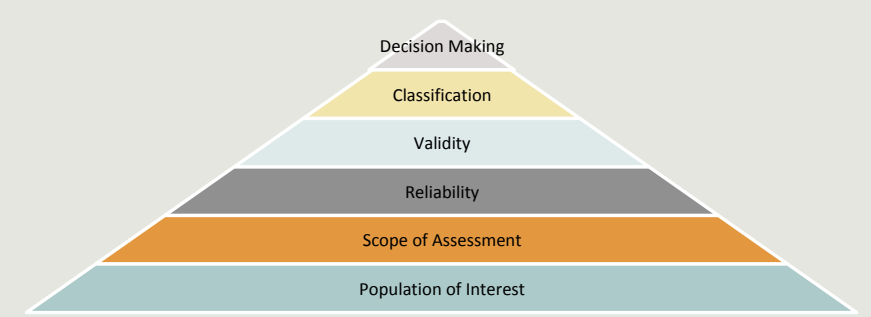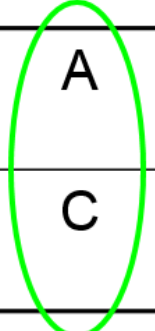
		Outcome	
		$<10^{th}$%ile	$>=10^{th}$%ile
Screener	At-Risk	A	B
	Not At-Risk	C	D

Sensitivity

Sensitivity: This indicates the percent of children who will be identified as having a problem **out of all the children who actually have a problem**.

Specificity

Specificity: This indicates the percent of children who will be identified as not having a problem **out of all the children who actually do not have a problem**.

Positive Predictive Power

PPP: This indicates the percent of children who will actually have a problem **out of all the children who were identified as having a problem**.

SAT10 Score

Fail   Pass

	Fail	Pass
Fail (SRI Score)	A	B
Pass	C	D

$$PPP = \frac{A}{A+B} = \frac{TP}{TP+FP}$$

Negative Predictive Power

NPP: This indicates the percent of children who will actually not have a problem **out of all the children who were identified as not having a problem**.

$$NPP = \frac{D}{C+D} = \frac{TN}{FN+TN}$$

Balancing index strength
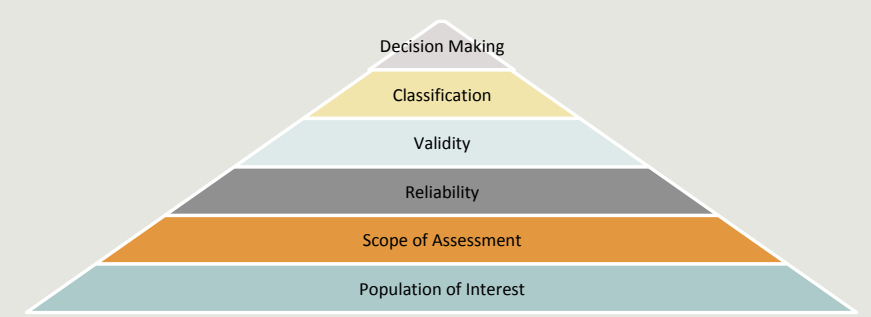
Increasing sensitivity means increasing the false positive rate

Increasing specificity means increasing the false negative rate

Increasing negative predictive power means increasing false positives

Increasing positive predictive power means increase false negatives

		SAT10 Score	
		Fail	Pass
SRI Score	Fail	A	B
	Pass	C	D

Translating into Action

Tools to Support Choosing and Using Screeners

National Center on Intensive Intervention

Evaluating: Rating Mechanism

Classification Accuracy

Convincing Evidence
- Lower Bound AUC Confidence Interval > .80, SE >= . 70, and SP >= .80 PLUS
- Q1 = Appropriate measure of academic performance used as an outcome (e.g., standardized word reading)
- Q2 = Risk was defined based on an RTI approach (e.g., 20th %ile) and with a consistent base rate
- Q3 = Classification analyses were done correctly

Partially Convincing Evidence
- Q1-Q3 answered Yes PLUS
- Either = Lower Bound AUC Confidence Interval .70-80 OR SE >= . 60, and SP >= .70

Unconvincing Evidence
- Does not meet Convincing or Partially Convincing

Convincing evidence   Partially convincing evidence   Unconvincing evidence

Reliability

Convincing
- Either = Model-based approach to reliability was used
- Or
  - Two types of reliability were reported
  - From **two** representative samples
  - Lower bound of confidence interval of reliability > .70

Partially Convincing
- Either = Model-based approach to reliability was used
- Or
  - Two types of reliability were reported
  - From **one** representative samples
  - Lower bound of confidence interval of reliability =.60-.70

Unconvincing

● Convincing evidence   ◐ Partially convincing evidence   ○ Unconvincing evidence

Validity

Convincing
- ◦ At least two types of justified analyses are reported
- ◦ Analyses are drawn from at least one representative sample of students across all ability levels
- ◦ Lower bound of the confidence interval > .60

Partially Convincing
- ◦ Analyses, measures, and sample were appropriate for two types of validity but evidence was mixed with not all lower bound of the confidence interval > .60

Unconvincing

Convincing evidence   Partially convincing evidence   Unconvincing evidence
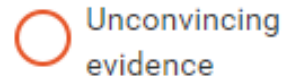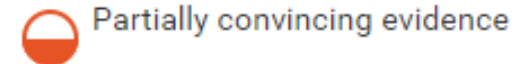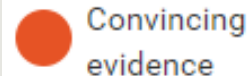
Other Information

Sample Representativeness
- National, Regional, Local
- With or without cross-validation

Bias Analysis
- Item-level or test-level

Administration Format

Administration and Scoring Time

Scoring Format

Types of Decision Rules

How do I get started?

- Gather a Team

- Determine Your Needs

- Determine Your Priorities

- Familiarize Yourself with the Content and Language

- Review the Data

- Ask Questions!

Helping you on the path

Population of Interest

1. How is the population defined?
   a. What is the intended age range for the assessment?
   b. How is the outcome (e.g., dyslexia, learning disability) defined?
2. When the screener was normed, is the sample reflective of the intended population?
   a. How similar is the norming sample to your local environment?
   b. Is the sample size for validating the screener sufficient for the analyses?
   c. Were multiple sites, states, or regions used to validate the screener?

Scope of Assessment

3. How is the outcome from question 1b operationally defined?
    a. What is the outcome by which students are judged to have a skill deficiency (e.g., standardized word reading test)?
    b. What cut-point is used on the outcome from question 3a to define "failure"?
    c. Is the cut-point from 3b reasonable for your local environment?
    d. Is the content on the screener reflective of what should be measured?
    e. Is the screener a measure of accuracy or automaticity?
        i. If the screener is computer adaptive, is the content developmentally appropriate for your local environment?
    f. Does the screener use more than one assessment?
        i. If yes, does the assessment provide guidance on how to use the scores in combination with each other?
        ii. If yes or no, does there appear to be good conceptual alignment between the screener and the outcome?

Statistical Considerations – Reliability

4. What type(s) of reliability are reported?
   a. If the screener is item-based, is internal consistency reported?
   b. If test-retest is reported, what is the spacing between testing occasions?
   c. If alternate-form or split-form reliability is reported, is another form of reliability reported?
   d. Are at least two forms of reliability reported?
      i. What level of reliability is reported?
   e. If the screener is not computer adaptive, is the reliability
      i. At least .80 (important for research decisions)?
      ii. At least .90 (important for clinical decisions)?
   f. If the screener is computer adaptive
      i. Is only marginal reliability reported (i.e., overall)?
      ii. Is reliability across a range of ability reported?
      iii. What is the level of reported reliability?

Statistical Considerations - Validity

5. Content Validity
    a. Has the domain been well defined (see question 1)?
    b. Is the domain relevant as defined
    c. Is the content appropriate for the local environment (see question 3.e.i)?
6. Substantive Validity
    a. Is there a reporting of how the test design matches the construct?
7. Structural Validity
    a. Are there tests of the factor structure/dimensionality reported (e.g., exploratory or confirmatory factor analysis)?
8. Generalizability
    a. For Bias, has one of the following types of analyses been used to test that the screener is not biased against subgroups (e.g., sex, race, poverty, students with disabilities, dual language learnings)
        i. Item-level bias analysis (e.g., differential item functioning)
        ii. Test-level bias analysis (e.g., differential classification accuracy)
9. External
    a. Convergent Validity
        i. Are correlations reported between the screener score and scores from an assessment on a related construct?
        ii. Are the correlations at least .60?
    b. Discriminant Validity
        i. Are correlations reported between the screener score and scores from an assessment on an unrelated construct?
        ii. Are the correlations no greater than .20?
    c. Predictive Validity
        i. Are correlations reported between the screener score at one time point and scores on an assessment at a later time point?
        ii. Are the correlations at least .20?
10. Consequential Validity
    a. Does the report document any intended or unintended side effects for those who are identified or misidentified based on the selected cut-points?

Statistical Considerations - Classification Accuracy

11. Is Sensitivity reported?
    a. Is it at least .80?
    b. Is a confidence interval reported and is the lower bound of the confidence interval at least .80?
12. Is Specificity reported?
    a. Is it at least .80?
    b. Is a confidence interval reported and is the lower bound of the confidence interval at least .80?
13. What is the Area under the curve?
    a. Is it at least .80?
    b. Is a confidence interval reported and is the lower bound of the confidence interval at least .80?
14. What is the False Positive rate?
15. What is the False Negative rate?

Some Resources

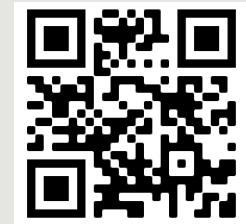National Center on Intensive Interventions
- https://charts.intensiveintervention.org/ascreening

National Center on Improving Literacy
◦ Understanding Screening
  ◦ Overall Screening and Assessment
  ◦ Reliability
  ◦ Validity
  ◦ Classification Accuracy
  ◦ Bias

Guiding Questions for Evaluating a Screener
◦ https://psyarxiv.com/vukt2/

Summary

Building a screener is a complex process

Choosing a screener is a complex decision

Evaluating the ingredients that make the sum is critical!

Use checklist starter when evaluating screener choices

Thank you!

ypetscher@fsu.edu
@yaacovp