

13th Panhellenic Conference of Psychological Research

Workshop on Test Development and Adaptation

Thomas Oakland
University of Florida
Oakland@ufl.edu

With appreciation to Kevin McGrew

Importance of Test Development

- “In an ever-changing world, psychological testing remains the flagship of applied psychology”
- Embretson (1996). The new rules of measurement. *Psychological Assessment*, 8, 341-349

Desired personal qualities

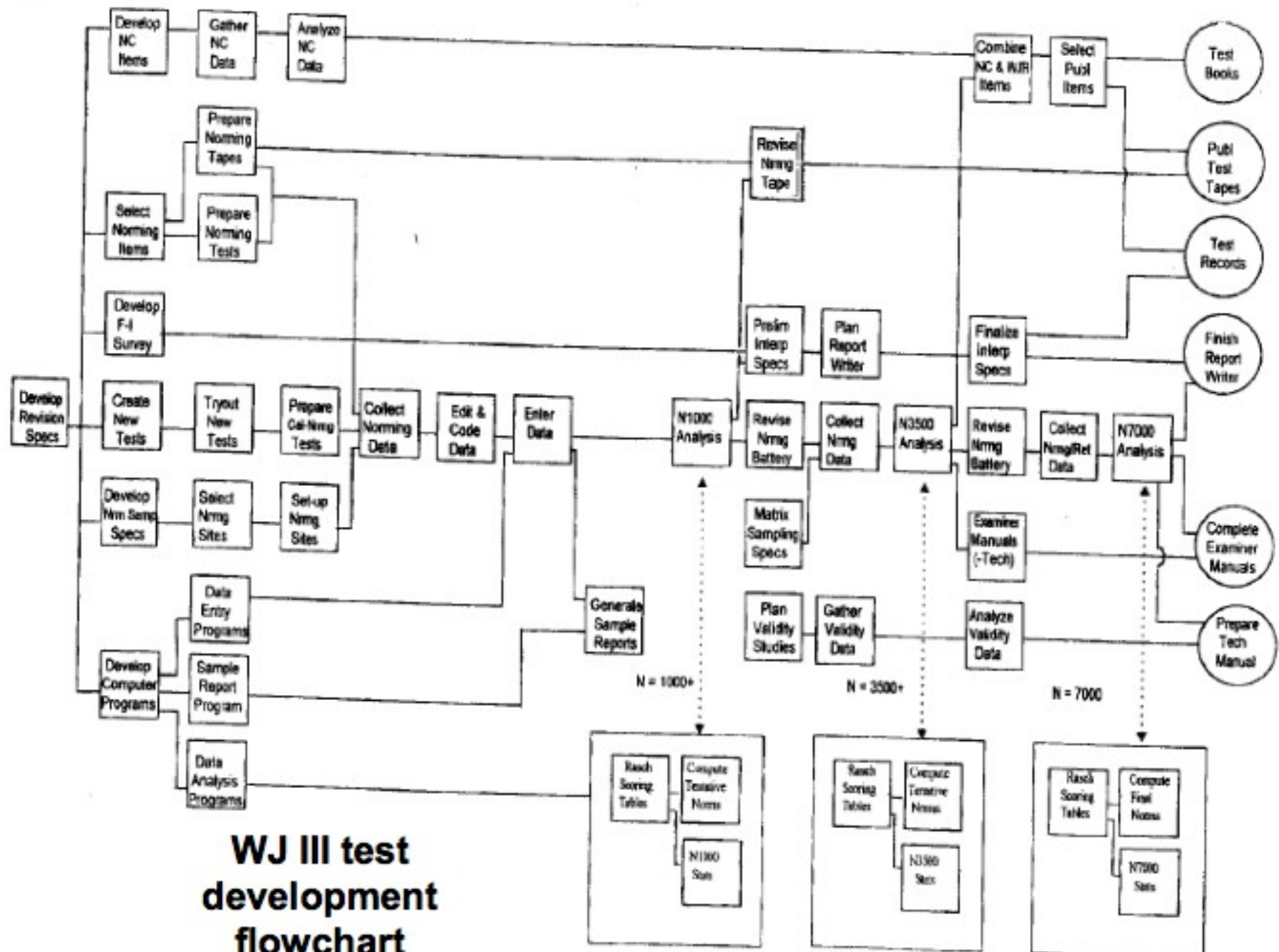
- Introverted = self-motivated
- A mix of attentiveness to detail and imaginativeness
- Intellectual curiosity
- Organized (obsessive-compulsive)
- Persistent
- Risk-taking and tough-skinned

Desired personal qualities: Troubles and Tips

- Commit years to this work
- Work in your area of specialization
- Consult with others who have needed expertise
- Form a test development team
- Do not expect to make a lot of money and recognize this may cost money

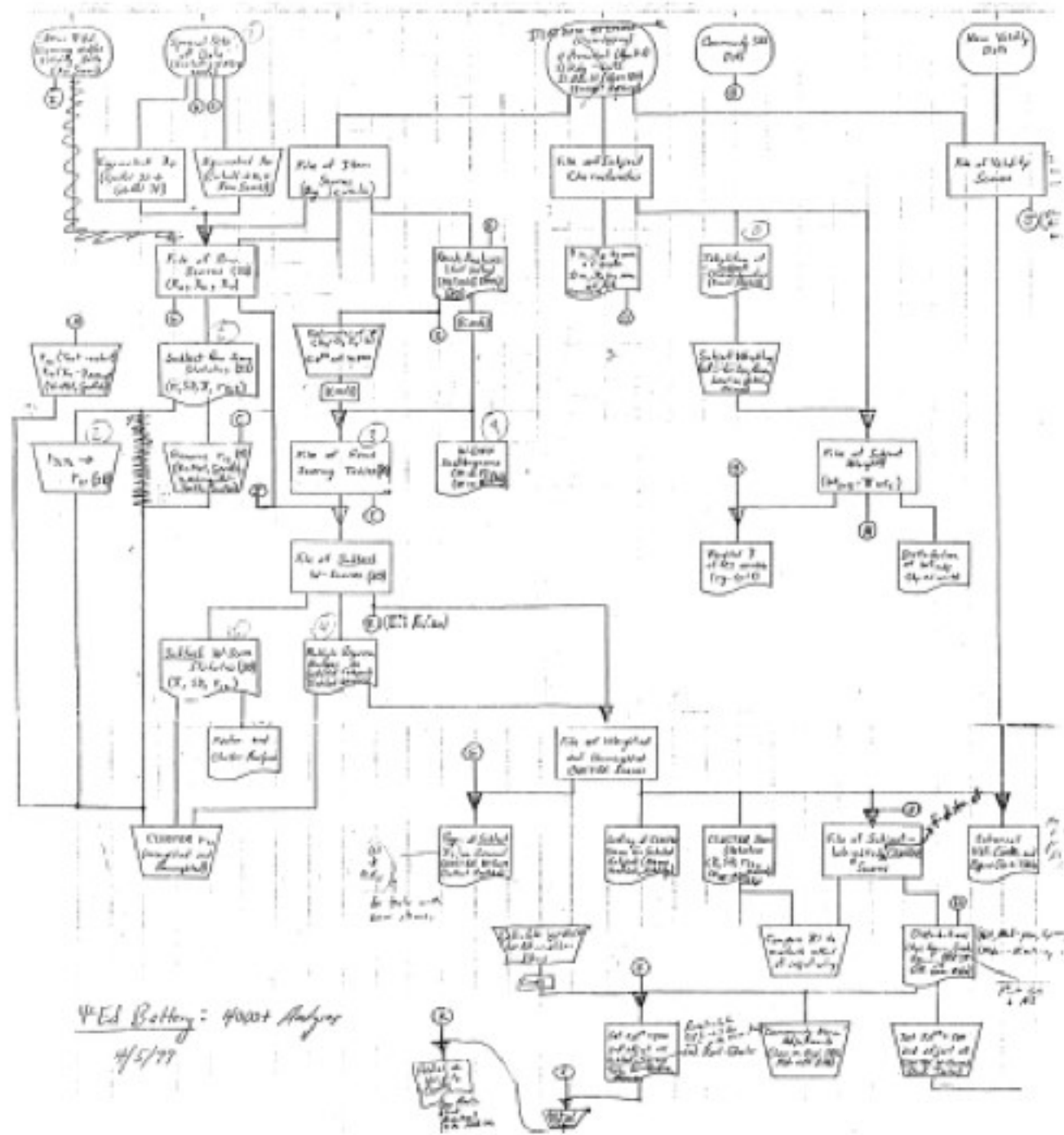
Desired personal qualities: Troubles and Tips

- Recognize the complexity of this work—
see next slides: one for test
development and another for data
analysis



**WJ III test
development
flowchart**

**1977 WJ
Data
analysis
plan
flowchart**



Core Ethical Principles

- Our contract with society
- Fidelity and Responsibility
 - Uphold professional standards of conduct
 - Clarify your professional roles and obligations
 - Accept appropriate responsibility
 - Manage conflicts of interest

Core Ethical Principles

- Justice
 - All people are entitled to benefit from services
 - Promote equal access to services
 - Minimize bias
 - Recognize the boundaries of one's competence

Core Ethical Principles

- Respect People's Rights and Dignity
 - Dignity and worth of all people
 - Right to
 - Privacy
 - Confidentiality
 - Self-determination
 - Diversity with respect to age, gender, gender identity, race, ethnicity, culture, national origin, religion, sexual orientation, disability, language, and SES

More Specific Ethical Test Standards

- Competence
- Delegating work to others
- Accuracy in interpreting test results
- Explaining test results
- Avoid assessment by unqualified persons
- Use of obsolete tests
- Maintaining test security
- Informed consent to develop tests
- **Plagiarism**

9 Categories of Persons Involved in Testing

- Test authors and those they supervise
- Test companies
- Test adaptors and those they supervise
- Those who educate others
- Those who purchase tests

9 Categories of Persons Involved in Testing

- Those who administer and score tests
- Those who make use of test data
- Those who are tested and their family
- Others who use test results (e.g. human resource personnel)

Where do you suggest we obtain needed scholarship?



Kit of Factor-Referenced Cognitive Tests

- A set of 72 tests that have been demonstrated to be consistent markers for 23 cognitive factors.

A research tool for studying:

- Reasoning
 - Verbal Ability
 - Spatial Ability
 - Memory
- and other cognitive processes

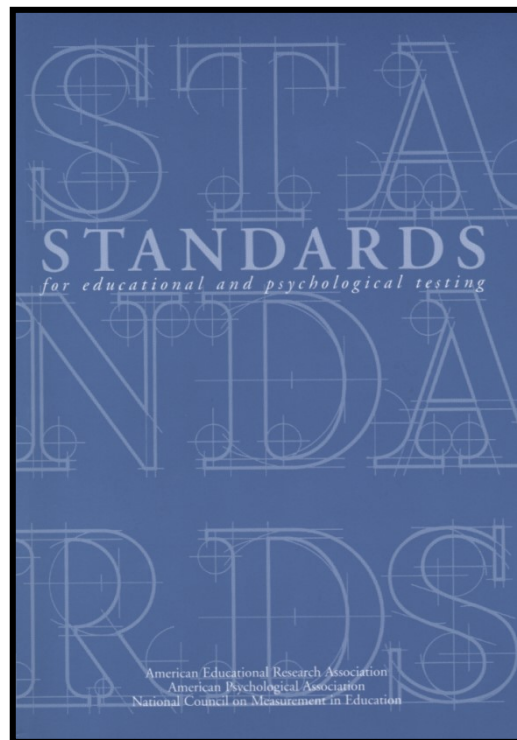
Important Information Sources for Clinical Tests

- Diagnostic and Statistical Manual of Mental Disorders and its international edition

Important Information Sources

- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: Author.

The bible of test development: The “Joint Standards”



Standards Chapters

I. Test Construction, Evaluation, and Documentation

Validity

Reliability

Test development and revision

Scales, norms, and score comparability

Test administration, scoring, and reporting

Supporting documentation for tests

Standards Chapters

II. Fairness in Testing

Fairness in testing and test use

Rights and responsibilities of test takers

Testing individuals of diverse linguistic
backgrounds

Testing individuals with disabilities

Standards Chapters

III. Testing Applications

Responsibilities of test users

Psychological testing and assessment

Educational testing and assessment

Testing in employment and credentialing

Testing in program evaluation and public
policy

Standards Chapter: Validity

The degree to which accumulated (empirical) evidence and theory support specific interpretations of test scores

Standards Chapter: Test Development

= the process of producing a measure of some aspect of an individual's characteristics by developing and combining items to form a test according to a specific plan. Consider...

- * test content

- * test format

- * context in which it will be used

Validation

Explicit statement of the proposed interpretation of test scores along with a rationale for the relevance of interpretation

Provide a scientifically sound validity statement to support the intended interpretations of tests scores.

Validation: Sources of Evidence

Test Content: to what extent is the test content consistent with the constructs and their domains being measured?

An example: 200 items subject to a review by experts in temperament →
100 taken to test standardization →
69 were selected as good items

Validation: Sources of Evidence

An analysis of the ways test-takers respond to the test items

Do the ways people respond actually assess the desired qualities?

Examples:

___ A test of reading comprehension GORT

___ Test-taking qualities that impede an accurate assessment (e.g. task involvement, mood, anxiety, distractibility, fatigue)

Validation: Sources of Evidence

Internal structure of the test: factor structure (exploratory and confirmatory factor analyses)

Relationships with other variables

___ Convergent evidence

___ Discriminant evidence

___ Predictive validity

The consequences of test use

Standards Chapter: Reliability

= consistency of data

Observed score = true score + error score

Thus, we work to reduce errors through
test

- * development
- * administration
- * scoring
- * interpretation (e.g. SEM)

Standards Chapter: Reliability

3 traditional estimates of reliability

- * the use of parallel forms of a test
- * test-retest
- * relationship between scores derived from individual items when a test is administered once (internal consistency: item-total correlations)

Standards Chapter: Bias

= when deficiencies in the test itself or the manner in which it is used result in different meanings for scores earned by members of different identifiable subgroups

Not mean score differences

Thus, bias occurs when the test's constructs or factors result in systematically different meanings across examinee subgroups.

Standards Chapter: Bias

Two key sources of potential bias:

___construct underrepresentation

___construct irrelevance

Other common sources of bias

___unclear directions

___does not meet the assumption of ~
exposure and thus the ability to
acquire the trait

___unsuitable norming

___language differences

10 Stages of Test Development

Substantive Stage

- 1. Identify the need for a test
- 2. Develop test specifications and design blueprint
- 3. Prepare initial item banks
- 4. Engage in small scale item tryouts

Stages of Test Development

Structural (Internal) & External Stages

- 5. Administer test to large and appropriate samples
- 6. Analyze and select items
- 7. Calculate norms and measurement statistics
- 8. Conduct reliability and validity studies and special analyses
- 9. Prepare finished test materials.
- 10. Market test

Identify a need: more than a good idea is needed

A test must be market-driven

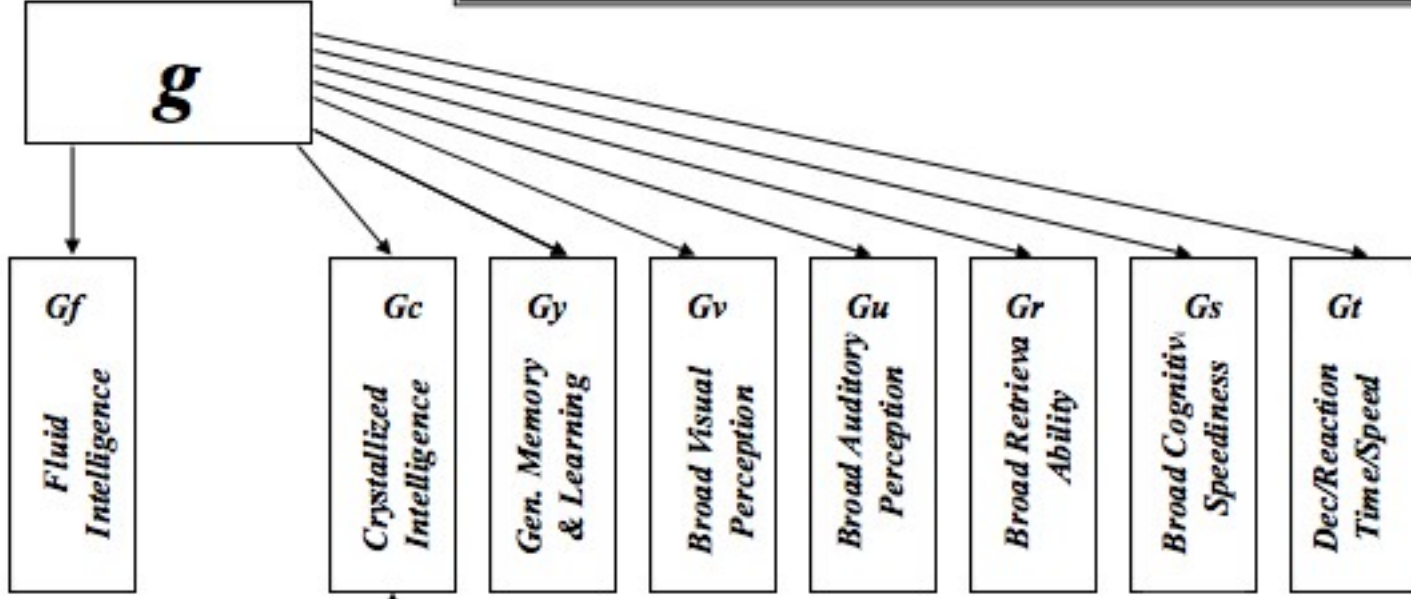
Hold focus group meetings

Botswana Project

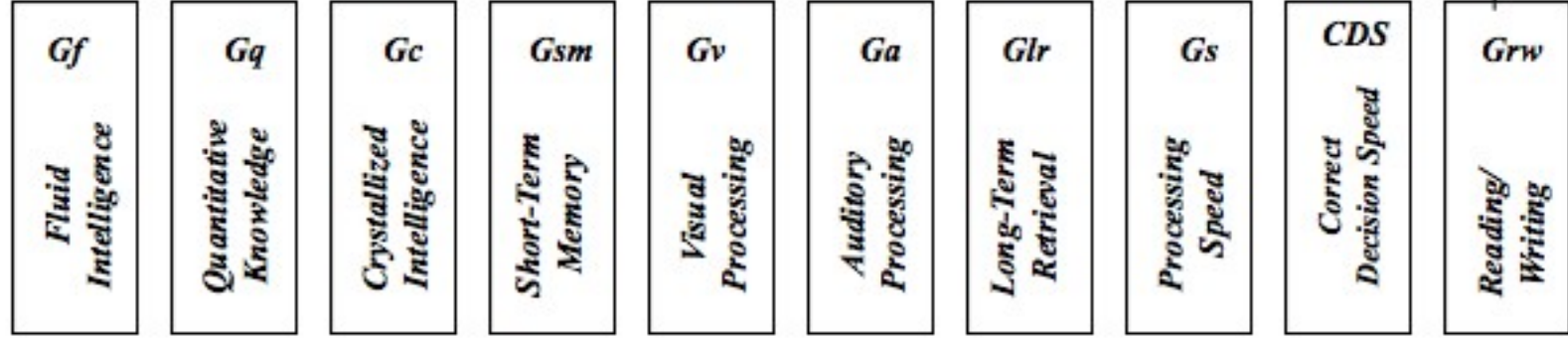
A change in theory (Wechsler's first definition of intelligence vs. CHC theory; adaptive behavior)

Carroll and Cattell-Horn Model Comparison

Carroll



Cattell-Horn



Identify a need: more than a good idea is needed

Changes in consumer preferences
(briefer tests)

Newly established national needs

Emerging clinical needs (Alzheimer,
autism)

Identify a need

How broad or general should a test be?

screening = a broad net (e.g. standardized interview)

general information (e.g. reading, general intelligence)

omnibus/comprehensive info (MMPI-II,
WJ > 40 subtests)

authentic assessment standards

8 qualities of authentic assessment

Acceptability: social worth

Authentic: natural methods and contexts

Collaboration: teamwork

Evidence: empirical foundation

Comprehensive (5 multi's)

Sensitivity: close measurement gradations

Universality: test applicable to all

Utility: Usefulness

Identify a need

What is your goal?

to diagnose (e.g., distinguish/discriminate)

to identify specific behaviors that that may
need clinical attention (thus, the items
must be very sensitive to clinical differences
(e.g., ASD)

to promote intervention and intervention
monitoring efforts

Identify a need: Troubles and Tips

Money-related issues are critical.

Most neophyte test developers start with a good idea that is not tested through a market analysis.

A test is like a child: it requires continued attention and **promotion**.

Is the market sufficiently large and financially able?

Develop Test Specifications & Blueprint

- A detailed description for the test
- 1st define the purpose of the test
 - Knowledge of the construct's theory
 - Define the construct operationally

Some Test Specifications & Blueprint Questions

Group or individually administered

Number of items needed in each
subtest (develop 2 to 3 Xs the final
number needed)

Item format, response mode, and
scoring rubrics

Desired psychometric properties of the
items and subtests

Develop Test Specifications & Blueprint

We will use the development of a measure of adaptive behavior as an example.

Test Specifications

adaptive behavior broadly defined

Practical everyday skills needed to **function and meet the demands of one's environment**, including the skills needed to **effectively and independently** take **care of oneself** and to **interact with others**

Test Specifications

The theory: based on a current definition from very reliable sources. Thus, what source is the most authoritative on adaptive behavior?

American Association of Mental Retardation.

What is its construct of adaptive behavior?

Test Specifications

A hierarchical model

10 adaptive skills embedded in
3 domains, and...

A general adaptive composite

Test Specifications

The ***Conceptual skill*** domain

Communication

Functional Academics

Self-Direction

The ***Social skill*** domain

Social Skills

Leisure

Test Specifications

The *Practical skill* domain

Community Use

Health and Safety

Home/School Living

Self-care

Test Specifications Questions

From what sources can I derive the items?

Are the items in each of the 10 skill areas (1) representative of the skill area yet (2) sufficiently different?

Do the items measure common, important, and observable behaviors?

Test Specifications Questions

Can items be written with sufficient clarity to ensure understanding? For example, ours are at the 8th grade reading level or lower (Sattler)

Do the 10 skill areas and their items **appear to be** consistent with the test's theoretical structure?

Test Specifications Include

Physical materials to be produced and their appearance

Test booklets

Manual(s)

Manipulatives

Audio/video tapes

Scoring keys

Computer scoring

Training materials

Report writing

Intervention planners Others

Test Specifications lead to...

Individually administered measures—
yet people could complete it on their
own

Respondents: who knew the person
well (e.g., parents, teachers,
self/adult)

Separate scales for 5-21, >21 and later
0-5

Test Specifications

Desired 25-30 items per scale

larger N for reliability and validity

smaller N for brevity

Decide on needed clinical and other special groups

Various DSM-groups depending on the test's market plan

Hungarians in Romania

Test Specifications

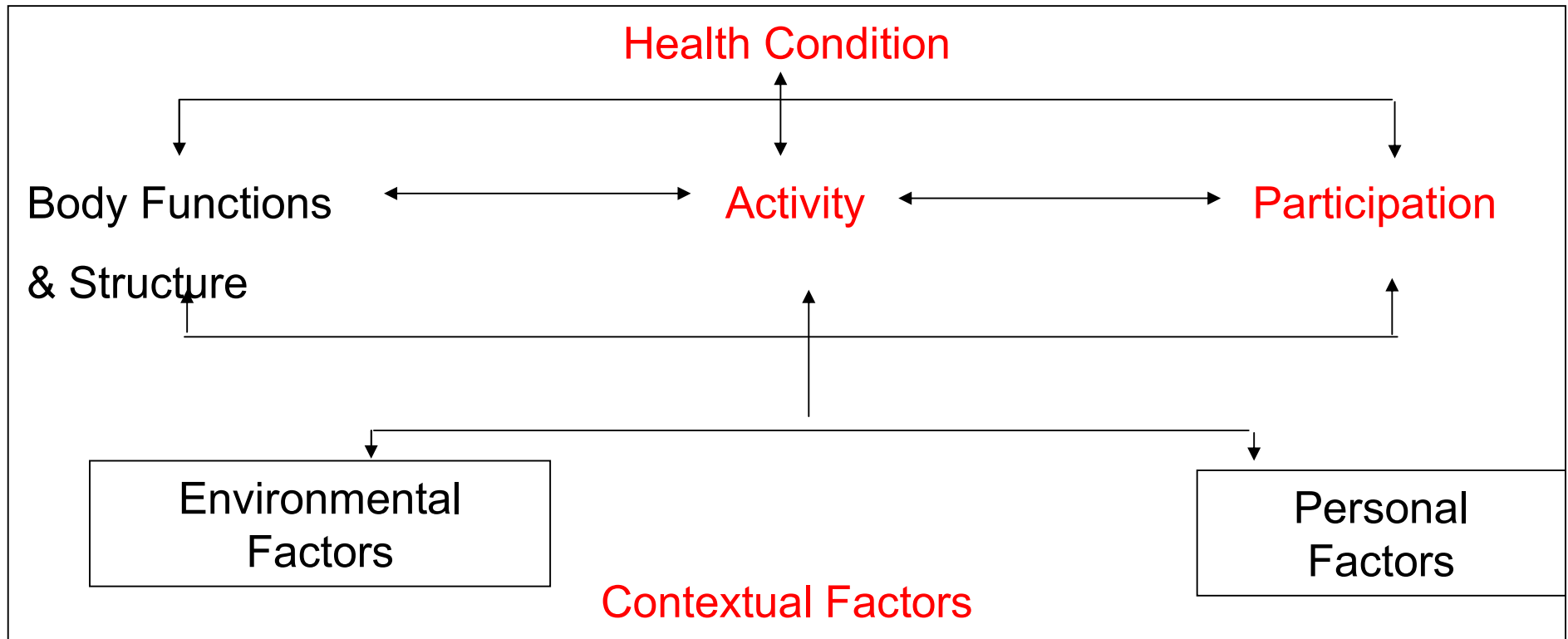
response mode

desired variability and thus used a 4
point scale

scoring rubrics (wanted to be
consistent with ICF model that
distinguished ability and performance)

ICF's Two Broad Parts

Health Conditions
Contextual Conditions



Test Specifications

Desired properties of the subtests & items

content validity

avoid

construct

underrepresentation

construct irrelevance

clear directions

meet the assumption of exposure

language initially in English and

now in various other

Random Sources of Error

Examinee's physical state

fatigue hunger

sickness alertness

visual and auditory acuity

motor coordination—especially fine
motor

Random Sources of Error

Examinee's psychological state

anxiety

memory

excitement

attentiveness

motivation

persistence

willingness to guess

not comprehending instructions

perceived importance of the test and
consequences of score

Random Sources of Error

External sources

timing brakes/interruptions

lighting acoustics

noise rapport

loss of standardization

Random Sources of Error

Test related

item format content sampling

scoring errors translations

inter-rater differences

luck

modes of administration

test reliability at different score levels

prior exposure to test content

Test Specifications: Troubles and Tips

The test's plan may be too grandiose.

Obtaining clinical data may be difficult

We never can anticipate all the problems. Thus, consider the test specifications to be a roadmap, not an unchangeable formula.

Prepare initial items

Prepare 2 to 3 Xs the desired number for each of the subtests

Subtests differ in difficulty for which to write items.

When developing the test over a broad age range, you may decide to not include some subtests.

Have all items reviewed by professionals who have different expertise

Prepare initial items

Consulted books and other scholarship

academic scholarship

developmental psychology

journal articles

source books

clinical scholarship (e.g., DSM)

Use your knowledge of behaviors

associated with the theory

Avoided copying other tests (avoid plagiarism)

Employ students and others

Our Item Preparation

Started with > 1000 items

Goal ~ 250 items per form

It may be time for coffee

Small-scale item tryouts

Divided the >1000 items into 5 test booklets of about 200 items each

Had about 200 people complete portions of each of the booklets

The respondents completed booklets for different ages

Thus, we had about 40 responses to each item

This lead to item statistics (described below)

Small-scale item tryouts

item statistics included data on...

approximate age level = difficulty level

response distribution statistics

number of guesses = item suitability

This → eliminating many items and writing more items

Item Tryouts: Troubles and Tips

- Writing items of moderate difficulty is easy. Writing easy and hard items is more difficult.
- Writing items for infants and young children is most difficult.
 - Behaviors often are subtle
 - Behaviors change quickly with age

Item Tryouts: Troubles and Tips

- Authors can be expected to fund item tryout activities.
- Expect to discover gaps in the data that require new items to be added.
- Item comprehension may be a problem and may need to be assessed through interviews
- Observe individually administered tests.

Item Tryouts: Troubles and Tips

- Some behaviors are not distributed normally, with fewer items at the upper end of the normal distribution
 - Medical qualities
 - Adaptive behavior
 - Psychopathology
- May need art work and thus an artist/designer

Item Tryouts: Troubles and Tips

- Test authors may need to go through more than one item tryout phase.
- Subsequent phases of test development and use will fail if the number of good items is not sufficient.
- Some items will be removed after all data are obtained.

Larger Scale Item Tryouts: Administer items to suitable samples

At this time you are acquiring data on a larger sample—however, not the larger national sample.

The goal is to acquire data for initial item statistics.

Many of the issues discussed previously apply here.

Administer items to suitable samples

Is this the stage at which you should attempt to obtain a contract?

More on this later.

Administer items to suitable samples

Samples are representative of the population with whom it will be used

Over-sample on some

high and low performing Ss

clinical groups

racial-ethnic and other minority groups

--anticipate the need to conduct validity studies with these data

Administer items to suitable samples

This work requires the physical materials to be prepared, printed, and stored

Test booklets Manual(s)

Manipulatives Audio/video tapes

Training materials

- On-site examiner training may be needed

Administer: Troubles and Tips

This step often is the most difficult: how to best obtain data economically, efficiently, and in a form readily codeable.

Will you compensate people if tested individually? If yes, how much?

Will you compensate institutions that allow you to obtain data? How much?

Other forms of compensation? (SSQ)

Administer: Troubles and Tips

You alone are unable to obtain national data. Thus, you need help from others.

*Friends: inexpensive yet unreliable.

*Colleagues: less expensive and may be motivated, especially with collecting clinical data. They may become allies of the test and thus help promote it.

Administer: Troubles and Tips

- *Students: rarely have contacts and are likely to be unmotivated.
- *Employ others
- *Contract with firms that acquire data
- *Establish a contract with a test development company

Analyze and Select Items

Now calibrate items. You may be norming the test items, not developing test norms.

Classical methods

discrimination

difficulty

distractibility

Analyze and Select Items

Classical methods

Items sample a domain. Use all items that contribute to a person's true score. The added correct items = one's observed score.

$$T_{\text{observed}} = T_{\text{true}} + T_{\text{error}}$$

Analyze and Select Items

Item Response Theory

Extends, not replaces, classical test theory (CTT).

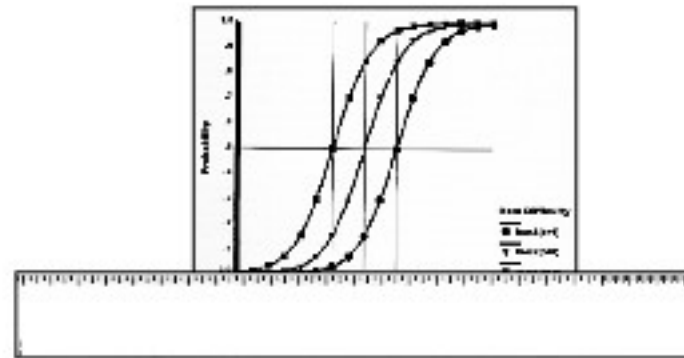
Based on probability theory

Determine item difficulty and arrange from low to high.

Based on a person's response patterns, what is the probability of a person passing a more difficult item?

See chapter 5 on IRT and CAT

Rasch IRT “norms” (calibrates) the scale !



Think of the items as now having been placed in **their proper position** on an **equal interval ruler or yardstick**....each item is a “tick” mark along the **latent trait scale**

Analyze and Select Items

Reorder, drop, add items

Possibly work on improving
item scoring keys

improving test instructions

Have an expert panel review items

Item Selection: Troubles and Tips

Adhere to the test design blueprint when selecting, deleting, and adding items.

Establish operational definitions of all constructs being measured.

Review and possibly use item formats that work well in other tests.

Remove items that don't work even if they are your favorites. (Dr. Wechsler)

Item Selection: Troubles and Tips

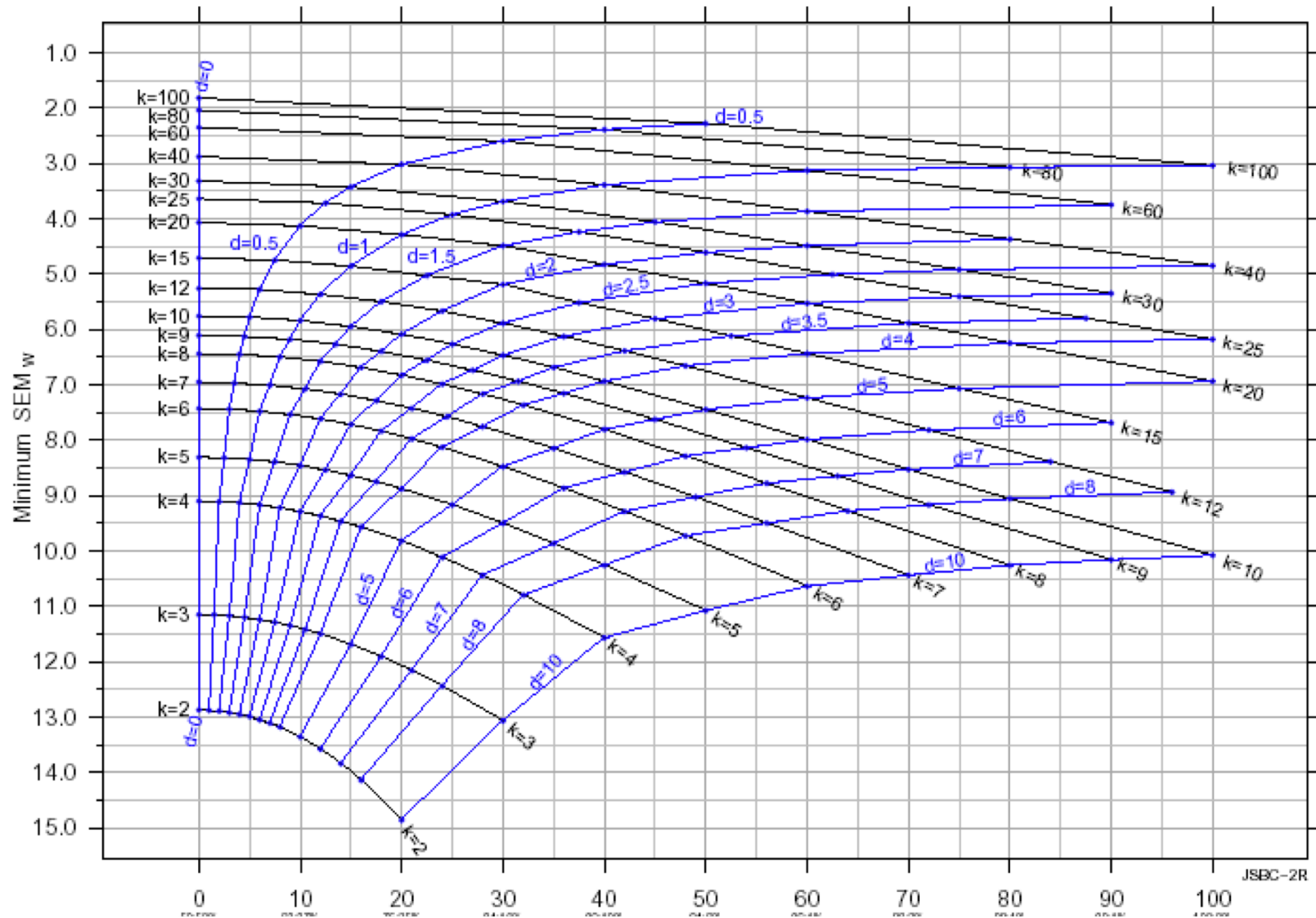
Use statistics you know. Avoid IRT/Rasch if they are unknown—or find a consultant.

Work to ensure some items are very easy and very difficult.

The test you norm should have 2 to 3 times the number of items desired on the final test.

Test Design Nomograph (W Scale)

R.W.Woodcock 1970, 1975, 1999, 2008



Item Selection: Troubles and Tips

Review test directions.

You may want to consider small-sample item tryouts.

As the test author, you must be intimately involved in this phase of the work.

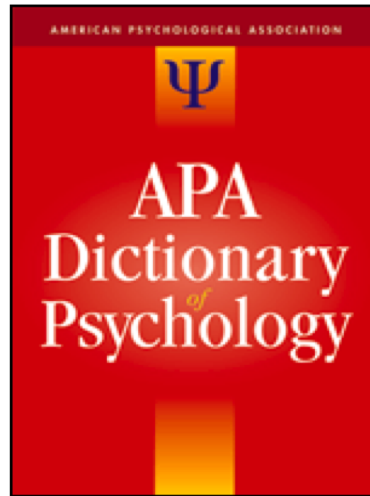
Obtain Norms

Do we always need norms?

Under what conditions may norms not be needed?

Obtain Norms If Needed

- Again, the goal is to obtain a nationally representative sample from which to develop test norms
- What size is needed?
High vs. low stakes tests: the higher the stakes, the larger the N



Norm: A standard or range of values that represents the typical performance of a group or of an individual (of a certain age, for example) against which comparisons can be made

Obtain Norms

- Again, the goal is to obtain a nationally representative sample from which to develop test norms
- Develop a standardization specification plan.
- Be sensitive to within country differences.
- Let's list some before going on.

Obtain Norms

- Possible Within country differences
- Age Race/ethnicity
- SES Religion
- Gender Length or residence
- Area lived in Education level
- Languages Other possible qualities

Obtain Norms

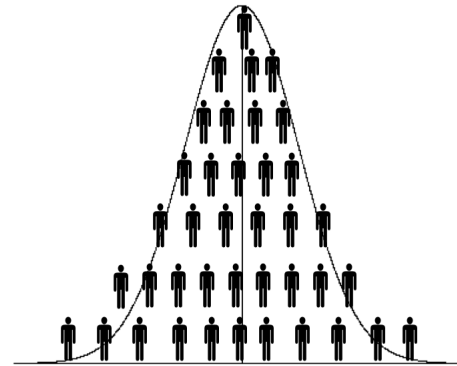
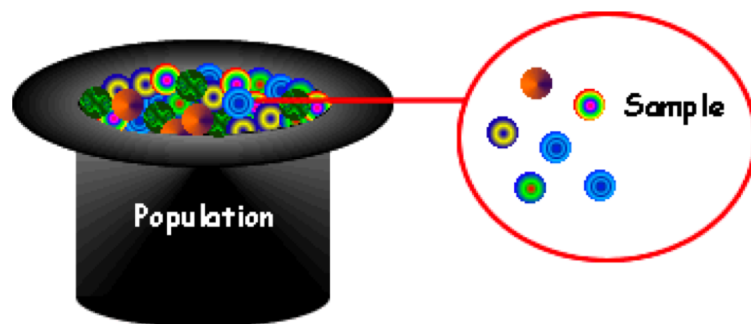
- On what criteria do you judge...?
- Age Race/ethnicity
- SES Religion
- Gender Length or residence
- Area lived in Education level
- Languages Other possible qualities

Obtain Norms

~/> 200 per age group for high stakes

Ns for group tests generally are larger than for individually administered tests.

Develop norm (standardization) sampling plan



The goal:
A **nationally representative** sample
from which to **develop test norms**

Obtain Norms

- Let's discuss the need to over-sample clinical groups.
- Who are possible clinical groups?
- How large should they be?
- Should they be representative of the general population?
- How can we gain access to them?

Obtain Norms

- Base sampling plan on best national census statistics, not one of convenience
- Establish quotas for the previously identified characteristics.
- Consider different sampling options: e.g.
 - Sample communities (especially on SES)
 - Then sample schools
 - Then sample students

Obtain Norms

- Determine the desired characteristics of examiners and then locate them
 - Availability Reliability
 - Experiences Willing to work for \$?
 - Age and gender Other qualities
 - Commitment to time
 - Willingness to travel
 - Attentiveness to detail

Obtain Norms

- Remember to over sample small groups, including clinical groups.
 - Why is this needed?
- Blanket (1 time) vs. continuous sample.
 - Enter and analyze the data continuously to detect possible difficulties.
 - If needed, stop data collection and make needed changes.

Obtain Norms

- Create a data coding system.
- Assign responsibilities for
 - The receipt of data
 - Inform examiners their data have been received and their quality
 - Contacting examiners to keep data flowing and the need for data from specific demographics
 - Data coding—double enter data
 - Data analyses

Obtain Norms: Troubles and Tips

SES: Do children from the same SES levels experience the similar environments and life styles?

large city vs. small rural community

large family vs. small family

Race: is race more important than SES?

Challenges with other demographic variables?

Obtain Norms: Troubles and Tips

Be mindful of the

___ sampling plan

___ need to acquire additional data

___ possible weighting of data

___ data needs for the next step: to determine the test's psychometric characteristics. If you consider it, include it. You later can discard it if not needed.

Obtain Norms: Troubles and Tips

Be mindful that examiners

___ may not be psychologists

___ may manufacture data

___ become lazy and not participate

___ may not record individually
administered data accurately

Consider conducting training, reliability checks, and close scrutiny to the first 3 to 4 protocols before payment.

Calculate Norms

Decide whether CTT, IRT, or a combination of methods will be used.

Examine expected group differences:

- Age Race/ethnicity
- SES Religion
- Gender Length or residence
- Area lived in Education levels
- Languages Other possible qualities

Calculate Norms

Decide on score values

Standard scores

$$M = 100 \quad SD = 15$$

$$M = 10 \quad SD = 3$$

T-Scores $M = 50 \quad SD = 10$

Age and grade equivalent scores

Others?

Calculate Norms

Examine expected group differences on the total test as well as its subtests.

Examine group differences in reference to their

means

standard deviations

skewness

Calculate Norms

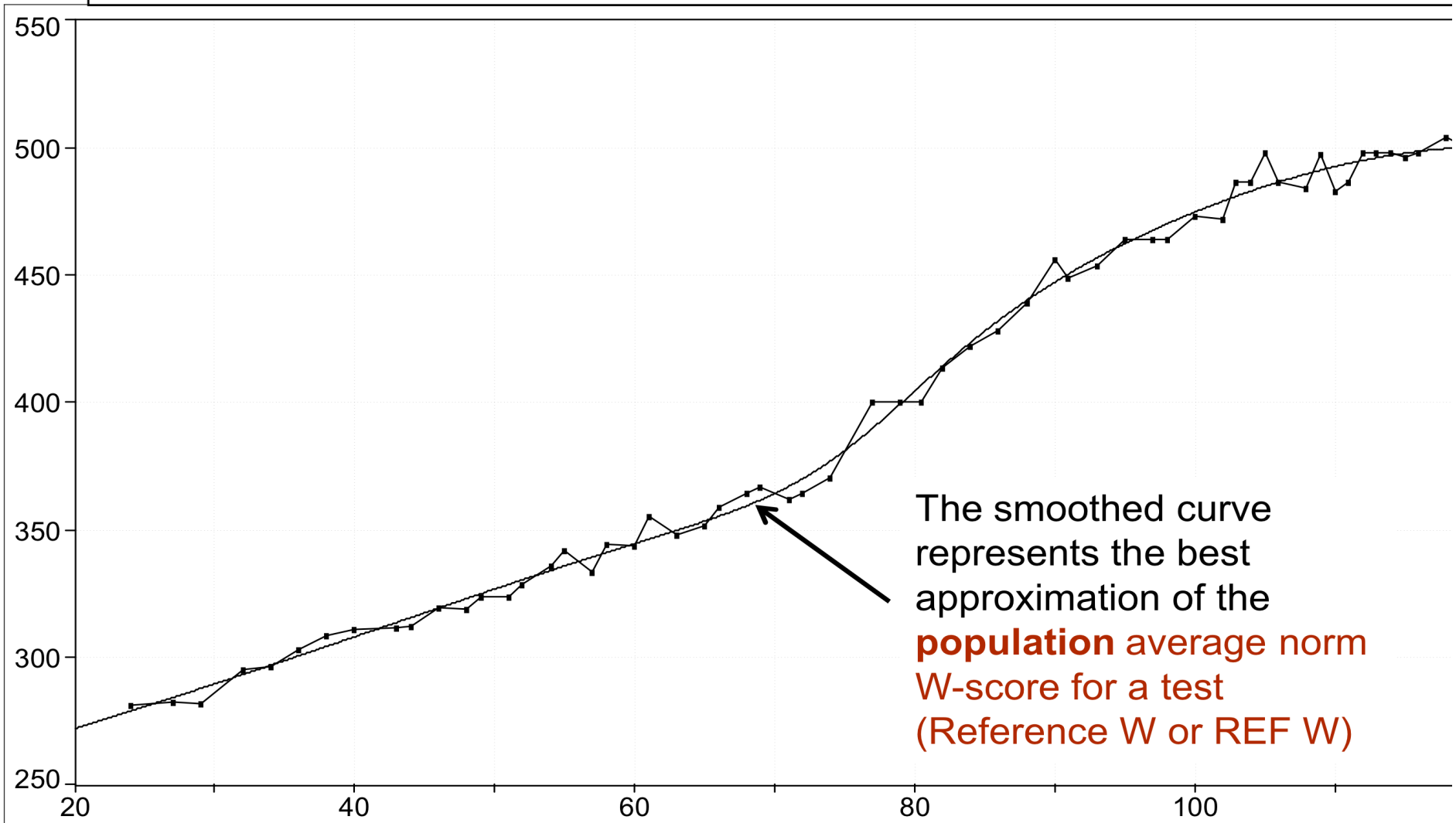
Differences can be expected:

SES: on tests of cognitive abilities

Gender: on tests of temperament and
the incidence of disorders

Others?

Letter-Word ID Ref W (20-120 months) polynomial curve generated solution
(using special curve fitting software)



Obtaining Developmental Scores (age/grade equivalents)

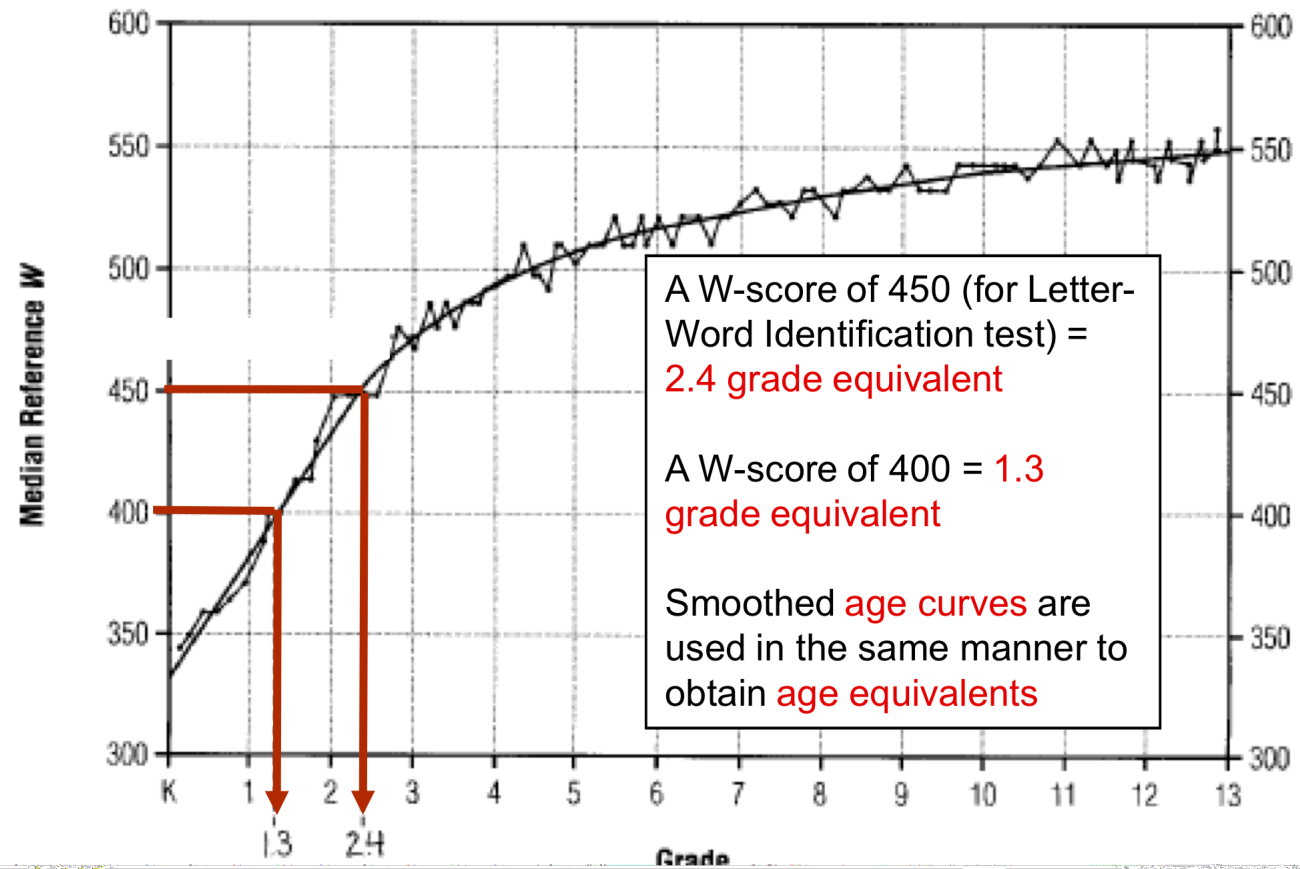


Figure 2-4.
Example of a smoothed curve
for WJ III ACH Test 1: Letter-
Word Identification

20

40

60

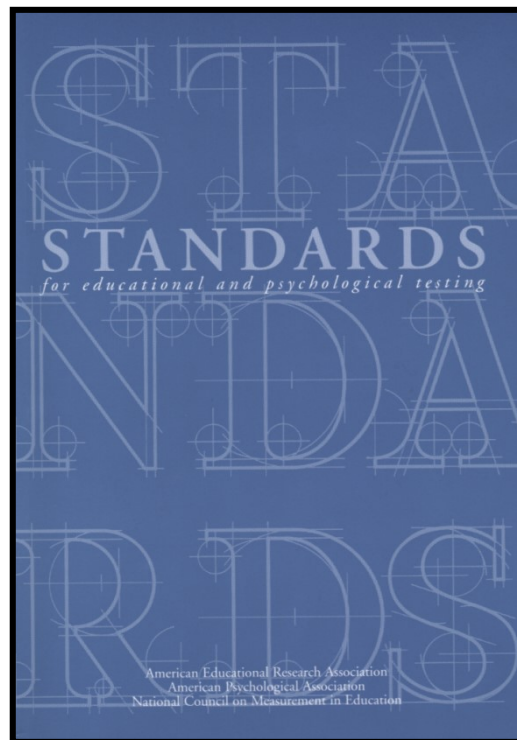
80

100

Calculate Norms

Unexpected group differences need attention. However, if the data were analyzed continuously as they were entered, unexpected differences should not be found toward the end of the norming process.

The bible of test development: The “Joint Standards”



Reliability

Reliability = the degree to which test scores for a group of test takers are consistent over repeated applications of a measurement procedure and hence are inferred to be dependable, and repeatable for an individual test taker; the degree to which scores are free of measurement error for a given group.

Reliability

Informal standards for reliability

- > 90 for individual decisions
- > 80 for group decisions
- > 70 for some research purposes
- < 70 probability of little value

Question: under what conditions may these standards be inappropriate?

Reliability Requires Variance: Troubles and Tips

Can we generalize reliability estimates based on a total group to subgroups?

What are the advantages and disadvantages of using...?

scores from

total group

ages/grades

demographic groups others?

Reliability Requires Variance: Troubles and Tips

Weaknesses may lead to
items added or dropped
subtests dropped
additional data acquired

Reliability Requires Variance: Troubles and Tips

Obtain external evaluations of your test as soon as possible, knowing the test must pass external standards before its release. Fix what can be fixed.

clinicians

psychometricians

specialists in the theory

Reliability Requires Variance: Troubles and Tips

A dilemma: professionals want shorter tests. Yet, shorter tests run the risk of displaying lower reliability and construct under-representation. What do we do?

Validity

Does a test have validity?

That is, how can we answer the question: Is X test valid?

Validity

The degree to which accumulated evidence and theory support specific **interpretations of test scores** entailed by the proposed **uses of tests**.

Responsibility for determining a test's validity rests with the test authors, publishers, and users.

Examine both internal and external forms of validity

Internal Forms of Validity

Examine the relationship between the test's content and construct it attempts to measure (e.g, content analysis)

On the ABAS-II,

1. are the subtests and domains seemingly consistent with theory?

2. are the items that assess work and other skills seemly

Internal Forms of Validity

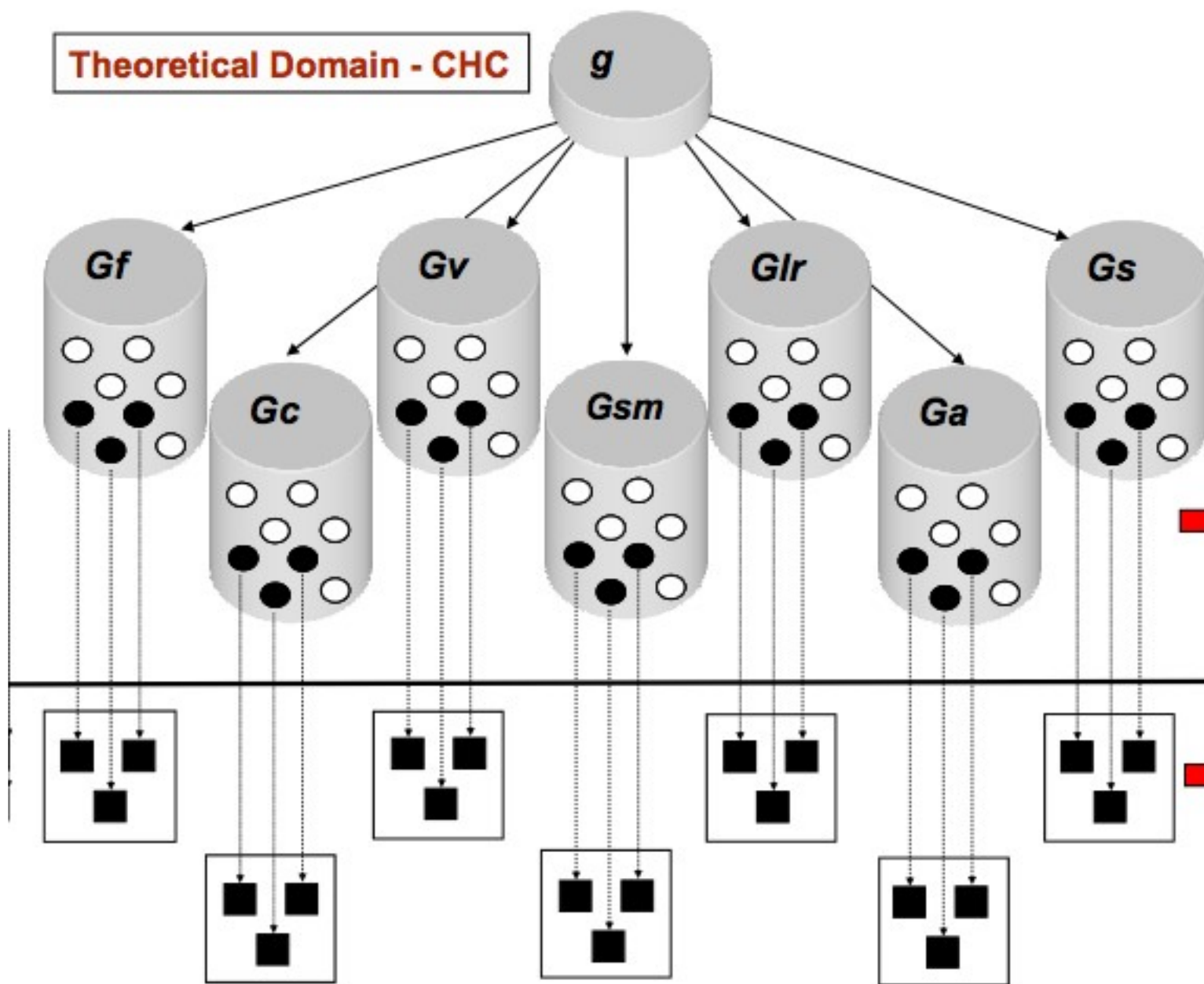
The relationship between the test's content and construct it attempts to measure through factor analysis

Exploratory

Confirmatory

Do both the domains and items cluster as expected?

Theoretical Domain - CHC



External evidence is focused on relations between test battery variables (measures or latent constructs) and **other external (outside of battery) constructs, measures, or criteria**

Measurement or empirical domain

Internal Forms of Validity

If we are developing a test to assess clinical qualities, do the items assess critical qualities?

Do the data show expected difference in reference to

- Age Race/ethnicity
- SES Religion
- Gender Length or residence
- Area lived in Education levels
- Languages Other possible qualities

Returning to Bias

Two key sources of potential bias:

___ construct underrepresentation

___ construct irrelevance

Other common sources of bias

___ unclear directions

___ does not meet the assumption of ~
exposure and thus the ability to
acquire the trait

___ language differences

External Forms of Validity

Let's consider what external evidence may be needed.

Support through

*convergent evidence (e.g.
correlates significantly
with other indices)

* discriminant evidence (e.g,
little to no correlation)

External Forms of Validity

Let's consider what external evidence may be needed.

Concurrent evidence

Predictive evidence

Other tests—especially those held in high esteem

Clinical differences (see following slide)

Results for Samples with Mental Retardation

Sample areas	GAC Mean	GAC \leq70 %	2+ skill 4 or below
MRMI (T, <i>n</i> =66)	73 (97)	50 (14)	76 (32)
MRMO (T, <i>n</i> =41)	59 (98)	70 (4)	100 (30)
Down's (T, <i>n</i> =22)	55 (100)	82 (5)	100 (23)

Note: Data for matched control group appears in parentheses.

Validity: Troubles and Tips

- What is the cost-effectiveness ratio?
- What if we can acquire only some external data?
- How can we overcome issues of privacy when collecting and recording data?
- How can we encourage others to engage in research on our test?

Validity: Troubles and Tips

- Assuming we have validity coefficients of ..., what evidence do they provide?
.20 .30 .40 .50 .60 .70
- What is the suitable size of validity quotients?
- Can we establish the same standards for validity coefficients that we do for reliability coefficients? Why or why not?

Validity: Troubles and Tips

- What may be some consequences of using the test—what on-going duties do we have?
 - My experiences with death row criminals

What may be some consequences of not using a test?

Validity: Troubles and Tips

What external forms of evidence will you rely on?

Consider developing more than one test concurrently.

Rely on existing and readily available clinical data—even if they are not perfect.

Finished Test Materials

Having completed the test's development, ...

How many copies do you make?

How much do you charge for purchase

workshops (paid or free?)

How do you market the test?

How do you get it in the hands of key persons who may promote the test?

Finished Test Materials

If the test has consumables,

do you have storage:

will they be readily

available in 5 years?

When may you consider revising the test? For example, how long is a test's shelf life?

Additional comments and questions you
may have

Keep in mind the importance of test development

- In an ever-changing world,
psychological testing remains
the flagship of applied
psychology

I have enjoyed my time with you

Best wishes on your continued success.