Test & Questionnaire Development and Item Writing

Invited presentation:
Monitoring Learning Achievement Workshop
Asmara, Eritrea,
5 December 2006

Anil Kanjee

Social science that makes a difference

HSRC
Human Sciences Research Council

Purpose

- To briefly review the process of:
  - test development and
  - item writing
- To remind participants about item writing rules
- All aspects will be addressed and covered in more detail in the working groups

National Education Quality Initiative

HSRC Team

- **Anil Kanjee - Questionnaires**
- **Gerda Diedricks - Maths**
- **Matthews Makghamatha – Languages**
- **Jenny Povey – Questionnaires**

National Education Quality Initiative

Overview: Test development Process

- Decide on purpose of assessment
- Develop table of specifications
- Item writing (and scoring)
- Item Review
- Assemble pilot instruments
- Trial testing
- Item analysis
- Item revision (discard)
- Assemble final instruments
- Develop administration manual
- Scoring and Reporting
- Questionnaire Development

National Education Quality Initiative

SRC
Sciences
ch Council

Overview: Test development Process

- **Decide on purpose of assessment**
- Develop table of specifications
- Item writing (and scoring)
- Item Review
- Assemble pilot instruments
- Trial testing
- Item analysis
- Item revision (discard)
- Assemble final instruments
- Develop administration manual
- Scoring and Reporting
- Questionnaire Development

National Education Quality Initiative

SRC
Sciences
ch Council

Defining the Purpose

- To obtain information levels of performance in
  - Grade 3 and 5
  - Grade 5
  - Mathematics, English, Mother Tongue
- To identify current conditions of learning and teaching in all primary schools
- To determine factors that affect learning for identifying relevant interventions to improve the functioning of the education system

National Education Quality Initiative

Overview: Test development Process

- Decide on purpose of assessment
- **Develop table of specifications**
- Item writing (and scoring)
- Item Review
- Assemble pilot instruments
- Trial testing
- Item analysis
- Item revision (discard)
- Assemble final instruments
- Develop administration manual
- Scoring and Reporting
- Questionnaire Development

National Education Quality Initiative

Table of Specification

- **Operational Definition: defining a construct in a way that is measurable. Sets boundaries on types of questions that will be asked.**

- **Decide on the number of items, subscales, domains of interest.**

Specify Administration and Scoring Procedures

- Administered in paper and pencil, computer, or oral format?
- Group or individual?
- Scored by test developer, taker, or user?

Overview: Test development Process

- Decide on purpose of assessment
- **Develop table of specifications**
  - **Cognitive level**
- Item writing (and scoring)
- Item Review
- Assemble pilot instruments
- Trial testing
- Item analysis
- Item revision (discard)
- Assemble final instruments
- Develop administration manual
- Scoring and Reporting
- Questionnaire Development

National Education Quality Initiative

SRC
Sciences
ch Council

Example – ToS – Maths

Content domains	Cognitive domains			
	Knowledge/ recall of information (40%)	Application of knowledge /solving routine problems (40%)	Analysis and problem solving/ Investigation (20%)	Total
Numbers, Operations and Quantitative reasoning (35%)	5	5	3	12
Patterns, Relationships and Algebraic thinking (14%)	3	3	2	9
Geometry and Partial reasoning (17%)	3	3	2	9
Measurements (20%)	3	3	2	9
Probability and Statistics (14%)	3	2	2	7
Total number of items	17	16	11	44

Blooms Taxonomy

- Level 1: Knowledge
- Level 2: Comprehension
- Level 3: Application
- Level 4: Analysis
- Level 5: Synthesis
- Level 6: Evaluation

National Education Quality Initiative

Knowledge

- At this level, one simply requires the recall of acquired knowledge. An assessment at this level can easily become a "Trivial Pursuit" exercise!
- Example:
  - What do these signs mean?

Comprehension

- At this level, knowledge of facts, theories, procedures etc. is assumed, and one tests for understanding of this knowledge.

- Example:
  - If you see this sign at 3:00 am what should you do?

Application

- Can someone apply the knowledge they have? Often requires observation.

- Example:
  - Can you stop at this sign?

Analysis

National Education Quality Initiative

- Analysis would include unstated assumptions, recognize logical fallacies, distinguish between facts and inferences, evaluate the relevancy of data, or analyze the organizational structure.

- Example:
  - You are a policeman and attend the scene of an accident. What facts should you gather?

Synthesis

- Using facts and experience present an argument, a well organized theme or speech, a creative short story (or poem or music), propose a plan, and integrate information into a plan.

- Example:
  - As a policeman you have gathered facts from the scene of an accident and you must now prepare a report that will be used in court.

Evaluation

- Pass judgment on, for example, the logical consistency of written material, the validity of experimental procedures or interpretation of data.
- Example:
  - You are the judge in a case to determine who is responsible for the accident

National Education Quality Initiative

Example – ToS – Maths

Content domains	Cognitive domains			
	Knowledge/ recall of information (40%)	Application of knowledge /solving routine problems (40%)	Analysis and problem solving/ Investigation (20%)	Total
Numbers, Operations and Quantitative reasoning (35%)	5	5	3	12
Patterns, Relationships and Algebraic thinking (14%)	3	3	2	9
Geometry and Partial reasoning (17%)	3	3	2	9
Measurements (20%)	3	3	2	9
Probability and Statistics (14%)	3	2	2	7
Total number of items	17	16	11	44

Overview: Test development Process

- Decide on purpose of assessment
- Develop table of specifications
- **Item writing (and scoring)**
- Item Review
- Assemble pilot instruments
- Trial testing
- Item analysis
- Item revision (discard)
- Assemble final instruments
- Develop administration manual
- Scoring and Reporting
- Questionnaire Development

National Education Quality Initiative

SRC
Sciences
ch Council

Paper & Pencil Item Types

- **Selected Response**
  - Multiple-Choice
  - Binary Choice: True-False; Right-Wrong
  - Matching
- **Constructed Response**
  - Short Answer
  - Completion
  - Essay

Writing Good Items

- Develop twice as many items as you plan to use in the final test—some items will be bad.
- Identify item topics by consulting the test plan for content validity
- Be sure each item is based on something central to the construct
- Items should not be things that anyone could guess.

Writing Good Items

- **Write each item in a clear, direct manner. Simple, short, with correct grammar and punctuation.**
- **Use vocabulary and language appropriate for the target population at the appropriate reading level. May need to use slang.**
- **Make all items independent: should not be able to figure out the answer on one item by the answer on another.**

Writing Good Items

- Each item should ask about only ONE thing—no double-barreled items.
- Should be concrete, clearly defined in the context of the question, with no jargon or acronyms.
- Response Options should match the question
- Ask a subject matter expert to review the test items to reduce ambiguity. There should be only one interpretation of each item.

National Education Quality Initiative

P/P Questions Selected Response

- Reading what learners choose
- **Positives**
  - Highly manageable
  - Easy to score
  - Generates multiple samples of domain
  - Raises mean achievement
- **Negatives**
  - Difficult to write quality questions
  - Tendency to surface level
  - Dependent on reading
  - Guessing factor
  - Recognition of answer

P/P Questions: Constructed Response SHORT

National Education Quality Initiative

- Reading what learners write in brief
- **Positives**
  - Students create answer to reveal knowledge
  - Easy to score
  - Easy to write
- **Negatives**
  - Scoring requires key
  - May be difficult to restrict range of possible correct answers
  - May be confounded with handwriting, spelling

LONG Constructed Response

- Reading the long answers students write
- **Positives**
  - Easy to write
  - Forces extended intellectual engagement with domain
- **Negatives**
  - Very hard to score consistently & reliably
  - Confounded with communication skills
  - Can be prepared in advance
  - Single sample of domain because of time factor

MULTIPLE CHOICE ITEMS

- **Use where the task calls for a single, clear answer to a question.**
- **When well designed, emphasize critical thinking and reasoning, rather than factual recall.**
- **Use when the range of possible correct answers is too broad – to focus thinking**
- **Use to remove load of writing; not thinking**

Reasons to use M-C items

- Easy to score
- Easy to sample widely across domain of interest
- Highly manageable
- Raise mean achievement – fewer missing data responses
- Students like them (Elley & Mangubhai, 1992)

National Education Quality Initiative

Disadvantages of M-C items

- Hard to write quality items
- Believed to test surface level processing, usually because of poor construction
- Guessing factor
- May require good reading level
- Recognition of answer & test wiseness

Anatomy of a multi-choice item

Stem

What is the best purpose of a test item?

Key

a. **To measure student learning**
b. **To keep the principal informed**
c. **To produce a mark in the markbook**
d. **To show how little has been learned**

Distractors

Writing M-C items: rules for stems

- Keep clear & concise – “specific is terrific!”
- Not too long to read
- Avoid negatively worded questions. Emphasise ***NOT*** if used
- Check the answer is not elsewhere in paper
- Avoid clues in grammar (*a/an; is/are etc*)
- Use interrogatives (*What is the name of this tool?)* or imperatives *(State the advantages of the …)* rather than sentence completion

Writing M-C items: rules for answers

National Education Quality Initiative

- **Only one correct answer – the key.**
- **Answer is actually correct.**
  Check, check, re-check!
- **Answer is sufficient to answer the question.**
- **No pattern of correct answers.**
- **Should not repeat words in stem.**
- **Use typical errors students make.**

Multiple Choice Distractors or *wrong answers*

- **Plausible**
  - not silly or plainly wrong
  - Connected to
    - a commonly held misunderstanding, or
    - An overgeneralisation or a narrowing of application
- **Similarity to each other and answer**
  - Similar length
  - Similar style as answer
  - Match the grammar or style of stem or question
- **Attract guessers & those with limited/weak knowledge**
- Arrange in a logical order – alphabetical, numerical, time series …
- Avoid qualifiers – e.g., ***sometimes***, ***never***, ***always***
- Avoid ***all of the above***, ***none of the above***

Generating distractors

- Common mis-understandings.

  Which is the value of 2+3? (*Could use the following)*

  a. -1 (2-3)
  b. 5 (correct)
  c. 6 (2x3)
  d. 8 (2^3)
  e. 23 (writing digits together as a single numeral)

- Mis-interpretations of text

- Present to students as open-ended item.

Number of response alternatives

- Typically three, four or five
- Four or five favoured over three – reduces guessing, increases discrimination
- Temptation to use *all of the above* and *none of the above* if using five.
- Four is most common
- Use three if entirely appropriate – *an acre is* ***larger/smaller/equal to*** *a hectare?*

National Education Quality Initiative

Example M-C Question

- *Which year is associated with the early European exploration of New Zealand?*
  - a) 1215
  - b) 1492
  - c) 1642
  - d) 1852

- **Reasons for options**:
  - a) Signing of Magna Carta; a significant date in British history – date may be known by students; has a 2 in it
  - b) Year European settlement of United States was begun; has a 2 in it
  - c) Year Abel Tasman first recorded sighting of New Zealand; has a 2 in it
  - d) Year New Zealand Parliament constituted; has a 2 in it

SR: Multiple Choice Variants

- Not just best of 4 or 5 labeled A to E.

- **Choose the best word for each sentence.**
  - He [went, gone] to the store.
  - The boat [sunk, sank] in the storm.
  - He looked for [accomodation, acommodation, accommodation, acomodation] at the hotel.

- **Circle the underlined letters to show where capitals are needed.**
  tomorrow is saturday and i am going to visit rangi.

Challenging Multiple Choice Questions

- Which of the following options is the best way to rewrite the underlined part of the sentence.

Since neither her nor the Dean were willing to veto the curriculum changes, they went into effect as of September 1.

(**A**). her nor the Dean were willing
(**B**). she nor the Dean was willing
(**C**). her nor the Dean wished
(**D**). she or the Dean was willing
(**E**). she nor the Dean were willing

- With which of the following statements would the Chief Meteorologist be most likely to agree?

(**A**). Forecasting is a dismal business
(**B**). Forecasters are generally pessimistic
(**C**). There is a growing demand for forecasters
(**D**). Forecasting is inspirational
(**E**). Forecasters are always right

Test of Objective Evidence

Each of the questions in the following set has a logical or “best” answer from its corresponding multiple choice answer set. Best answer means the answer has the highest probability of being the correct one in accordance with the information at your disposal. There is no particular clue in the spelling of the words and there are no hidden meanings. Please record your eight answers.

Questions 1--2

1. ***The purpose of the cluss infurmpaling is to remove***
   a. cluss-prags
   b. tremails
   c. cloughs
   d. pluomots ( )

2. ***Trassig is true when***
   a. clump trasses the von
   b. the viskal flans, if the viskal is donwil or zortil
   c. the belgo fruls
   d. dissels lisk easily ( )

Questions 3--4

3. ***The sigia frequently overfesks the trelsum because***

a. all sigias are mellious
b. all sigias are always votial
c. the trelsum is usually tarious
d. no trelsa are feskable ( )

4. ***Which of the following is always present when trossels are being gruven?***

a. rint and yost
b. Yost
c. shum and Yost
d. yost and plone ( )

Instructions for Test-taker

- Administrator usually read script to test taker.
- Also appears in writing on the test
- Should be thorough but concise.

Writing Administration Instructions

- Must have detailed instructions on how to give the test to ensure standard administration
- Instructions should cover:
  - Group vs. individual administration?
  - Requirements for location, such as privacy, quiety, comfortable chairs, tables, or desks.
  - Required equipment (stopwatches, #2 pencil, etc)
  - Time limitations, or average adminstration time
  - Introduction scripts
  - Instructions on how to respond to test-taker questions.

Overview: Test development Process

- Decide on purpose of assessment
- Develop table of specifications
- Item writing (and scoring)
- Item Review
- **Assemble pilot instruments**
- **Trial testing**
- Item analysis
- Item revision (discard)
- Assemble final instruments
- Develop administration manual
- Scoring and Reporting
- Questionnaire Development

National Education Quality Initiative

SRC
Sciences
ch Council

Pilot test

- A scientific investigation of a new test's reliability and validity for its specified purpose.
- Give test to sample of interest.
- Get qualitative and quantitative feedback.

Overview: Test development Process

- Decide on purpose of assessment
- Develop table of specifications
- Item writing (and scoring)
- Item Review
- Assemble pilot instruments
- Trial testing
- **Item analysis**
- Item revision (discard)
- Assemble final instruments
- Develop administration manual
- Scoring and Reporting
- Questionnaire Development

National Education Quality Initiative

SRC
Sciences
ch Council

Quantitative Item Analysis

- **Item Difficulty**
  - **Percent that respond correctly to an item (p value)**
  - **For personality test, percent that say “yes” to the item.**
  - **Want those items that differentiate people - no 1.0 p values**
  - **Ideally want p values of .5, or 50%**
  - **Extremes are discarded (e.g. <.2, >.8)**

Item discrimination

- How well does a particular item discriminate between high and low scorers on whole test
- Divide test takers into thirds
- D = #of high scorers who answered the question correctly minus #of low scorers that answered it correctly.
- E.g., Question 1: 80% in the top third got it right, and 20% in the bottom third got it right. D=60%
- Want D to be high and positive.

Inter-item correlations

- Inter-item correlation matrix: shows correlation of every item with every other item.
- Low inter-item corr means items is probably not measuring the same thing as the other items
- Retain those that are most highly related to a criterion.
- Empirically keyed tests: developed by examining each item's correlation with a criterion.

Item Characteristic Curves

- Based on item-response theory
- A graphical of probability of an item being answered correctly given the individual's level of ability on the construct being measured.
- Graphs combine information about the items difficulty and discrimination.

Qualitative Item Analysis

- Can use questionnaires for test takers to elicit feedback on various characteristics of the items (cultural sensitivity, face validity, test fairness, language, etc.)
- Can also get feedback from expert panels.

Overview: Test development Process

- Decide on purpose of assessment
- Develop table of specifications
- Item writing (and scoring)
- Item Review
- Assemble pilot instruments
- Trial testing
- Item analysis
- **Item revision (discard)**
- Assemble final instruments
- Develop administration manual
- Scoring and Reporting
- Questionnaire Development

National Education Quality Initiative

Revising the Test

- **Can use all of this information to choose your best items.**
- **Once best items are chosen, they should be tested again.**

Validation and Cross-Validation

- After final items chosen, you should begin to collect validity data on your test.
- Findings should be replicated on more that one sample, as you may find “shrinkage.”

Overview: Test development Process

National Education Quality Initiative

- Decide on purpose of assessment
- Develop table of specifications
- Item writing (and scoring)
- Item Review
- Assemble pilot instruments
- Trial testing
- Item analysis
- Item revision (discard)
- Assemble final instruments
- **Develop administration manual**
- Scoring and Reporting
- Questionnaire Development

Developing the Test Manual

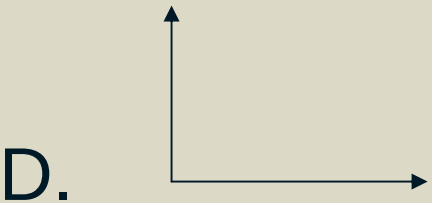
- **Should include:**
  - **Rationale for constructing the test**
  - **A history of the developmental process**
  - **Results of validation studies**
  - **Appropriate target audience**
  - **Instructions for administration**
  - **Norms**
  - **Information on Score interpretation**

Sample Original Item

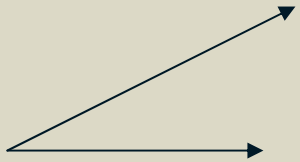
Ramón is building a doghouse. He wants the roof of the doghouse to be at an angle that is more than 90° but less than 110°. Which angle below could he use for the roof?

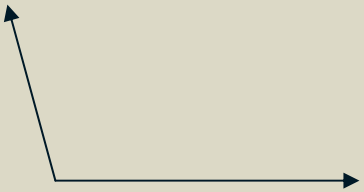
A.

B.

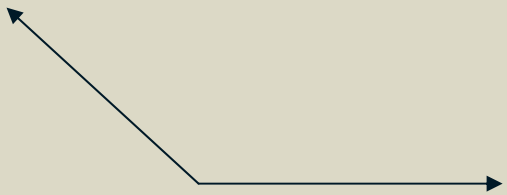


C.



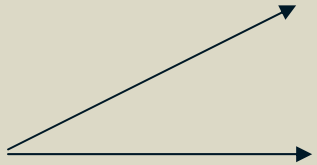


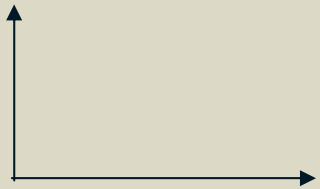
D.

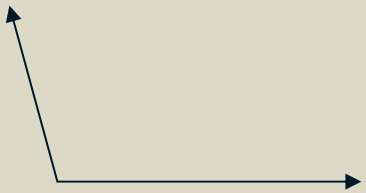


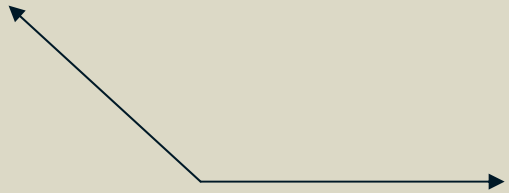
Revised Item

Which angle is more than 90° and less than 110°?

A. 

B. 

C. 

D. 

What changed?

Design element #2: Construct more precisely defined.

Design element #3: Bias eliminated (dog house, Ramón)

Design element #4: "Built in accommodations" – un-timed, students circled answer on paper, did not bubble

Design element #5: Simple instructions and procedures

Design element #6: More comprehensible language

Design element #7: Larger font

Overview: Test development Process

- Decide on purpose of assessment
- Develop table of specifications
- Item writing (and scoring)
- Item Review
- Assemble pilot instruments
- Trial testing
- Item analysis
- Item revision (discard)
- Assemble final instruments
- Develop administration manual
- **Scoring and Reporting**
- Questionnaire Development

National Education Quality Initiative

Scoring/Reporting - Reading

- ***Level 1 -- Non-reader*****:** A person who fails to identify at least seven out of eight given individual letters.
- ***Level 2 -- Rudimentary-level reader*****:** A person who correctly identifies a minimum of seven out of eight given individual letters.
- ***Level 3 -- Beginning reader*****:** A person who correctly completes six out of eight tasks requiring matching a simple word to a picture, and identifying simple words.
- ***Level 4 -- Minimally competent reader*****:** A person who correctly answers three out of five literal comprehension questions based on a passage.

Scoring/Reporting - Writer

- ***Level 1 -- Non-writer*****:** A person who fails to write at least six out of eight named letters.
- ***Level 2 -- Rudimentary-level writer*****:** A person who correctly writes at least six of eight named letters.
- ***Level 3 -- Beginning writer*****:** A person who writes at least four out of six dictated simple words and sentences which can be identified (allowing for errors).
- ***Level 4 -- Minimally competent writer*****:** A person who writes a short (12-word minimum) understandable text (allowing for errors in spelling and grammar) within a 6 minute period when given a pictorial prompt.

Scoring/Reporting – Maths

- ***Level 1 -- Non-numerate (written)*****:** A person who fails to correctly answer at least six out eight tasks involving recognition of four one-digit and two-digit numbers, writing two one-digit numbers, and recognizing two simple geometric shapes.
- ***Level 2-- Rudimentary-level numerate (written)*****:** A person who correctly answers at least six out eight tasks involving recognition of four one-digit and two-digit numbers, writing two one-digit numbers, and recognizing of two simple geometric shapes.
- ***Level 3 -- Low-level numerate (written)*****:** A person who correctly answers at least three out of four simple computations involving the four basic mathematical functions (addition, subtraction, multiplication, division).
- ***Level 4 -- Minimally-competent numerate (written)*****:** A person who correctly answers at least three out of four problems involving the four basic mathematical functions.

Overview: Test development Process

- Decide on purpose of assessment
- Develop table of specifications
- Item writing (and scoring)
- Item Review
- Assemble pilot instruments
- Trial testing
- Item analysis
- Item revision (discard)
- Assemble final instruments
- Develop administration manual
- Scoring and Reporting
- **Questionnaire Development**

National Education Quality Initiative

Questionnaire Development

Interviews and Questionnaires

National Education Quality Initiative

- How to write good questions?
  - Interview questions less rigid than surveys
  - Tradeoff: Interviews:
    - Take more time
    - Cover fewer points of view
    - Get richer information
    - Useful for formulating survey questions
    -
- How to design questions?
  - Questionnaire Design
  - Person-to-Person data collection
  - Focus groups
  - Web Surveys

Designing Questions

- The information obtained by each question will be specific to the information you will need in your analysis.
- Therefore, before you compose any questions:
  - Think through your research questions and objectives
  - Think about how you will conduct your analysis of the results

National Education Quality Initiative

Do:

- **Use simple wording**
- **Be brief**
- **Be specific**

National Education Quality Initiative

Social science that makes a difference

Slide adapted from Penn State Survey Research Center

HSRC Human Sciences Research Council

Do not:

- Be vague
- Be condescending or talk down to respondent
- Use biased wording
- Use abbreviations or scientific jargon
- Be redundant

National Education Quality Initiative

Question Writing Principles

- Scales are always relative to respondents' experience
- Scale should allow for maximum variability
- Use a balanced scale
- Be careful about responses of 'neutral' or 'no opinion' versus 'don't know'
- Use item-in-a-series response categories carefully

National Education Quality Initiative

Question-Writing Principles

- **Questions should ask for only 1 piece of information, so avoid:**
  - **Asking two questions at once**
  - **Asking questions that contain assumptions**
  - **Asking questions that have hidden contingencies**

Question Writing Principles

- **Question wording should ensure that every respondent will be answering the same thing, so avoid:**
  - **Ambiguous wording or wording that means different things to different respondents**
  - **Using terms for which the definition can vary. (If it is unavoidable, provide the respondent with a definition.)**
  - **Being ambiguous about the time period the respondent should consider**
  - **Asking complex questions (double-barreled)**

Question order

- Questions should be ordered so as to seem logical to the respondent
- First questions should be relevant and easy
- Questions are effectively ordered from most salient to least salient
- Demographic questions should not be covered at the beginning
- Potentially objectionable questions are placed near the end

Questionnaire design problems

- Some possible threats to accuracy:
  - Questions not being understood as intended
  - Not adequately capturing respondent's experience
  - Posing a challenging response task
- Problems may not be visible in the actual survey data
- How can we find these before data collection?

Social Desirability

- A pattern of response in which the respondent answers items in a way that presents themselves in a positive light.
  - Can include a scale that detects these people so social desirability can be statistically controlled.
  - Examining the correlation between a measure of social desirability and test items may reveal bad items.
  - Instruct test-takers on importance of honesty.

Acquiescence

- **Tendency to agree with whatever statement is offered, no matter what.**
  - **Are you happy? Yes. Are you sad? Yes.**
  - **Control by “reversing” a proportion of the items.**

Random Responding

- **Answer randomly without reading or understanding the items.**
  - **Can be due to lack of motivation, low reading level, neurocognitive, attentional problems**
  - **Can detect by having two questions asking the same thing at different points in the test—should be answered the same way.**
  - **Give items that all people will generally answer the same way: T/F: Child abuse is bad.**

Faking

- Also called malingering
- Subtle questions may be able to detect faking.
  - Does subject report symptoms that seem like they would be a symptom of the disorder, but in fact are rare in disordered people? Do they endorse the less-well known symptoms of the disorder?
  - Give easy questions.