COLLEGE OF EDUCATION AND HUMAN DEVELOPMENT | UNIVERSITY OF MINNESOTA

Michael C. Rodriguez

Measurement Essentials that Support *Learning*

Quantitative Methods in Education
Department of Educational Psychology

Measurement

Enduring Themes:

What to measure
How to measure

UNIVERSITY OF MINNESOTA | COLLEGE OF EDUCATION AND HUMAN DEVELOPMENT

Robert Ebel

Item-writing is an art. It requires an uncommon combination of special abilities. It is mastered only through extensive and critically supervised practice…. Item-writing is essentially creative. Each item as it is being written presents new problems and new opportunities.

(Ebel, 1951, p. 185)

Robert Ebel

Each item as it is being written presents new problems and new opportunities. Just as there can be no set formulas for producing a good story or a good painting, so there can be no set of rules that will guarantee the production of good test items. Principles can be established and suggestions offered, but it is the item writer's judgment in the application (and occasional disregard) of these principles and suggestions that determines whether good items or mediocre ones will be produced.

(Ebel, 1951, p. 185)

Mark Reckase

- *Test items are complicated. They are the equivalent of small poems.*
- *Credit should be given for great item writing.*
- *Cognitive scientists have long identified multiple cognitive skills that are required to interact with an item (e.g., reading, specialized vocabulary, knowledge of item formats, knowledge of the subject area).*

Mark Reckase

- *A careful review of any testing program will identify poorly worded test items... written by persons with minimal training and inadequate insights into their audience.*
- *We need to do much more work to produce quality test items.*

2009 NCME Presidential Address

From Art to Science

- Item writing is an art and a science
- Item writing requires supervised training
- Empirical research on item writing began in 1920s
- Multiple-choice items are capable of measuring a wide range of content and cognitive domains in a short time with a high level of reliability
- Not all students are able to perform on today's multiple-choice tests without accommodations or modifications

Measurement Essentials

New attention in the measurement community is on building

ASSESSMENT ***FOR*** LEARNING

in all forms of assessment

including large scale assessment and classroom assessment

Evidence Centered Design

- **Articulate Student Learning Outcomes (SLOs) and Claims**
  "What do we want to say about our students?"
- **Identify Evidence to Support Claims**
  "What can our students do to demonstrate the knowledge, skills, and abilities that are claimed by the department?"
- **Develop Assessments to gather Evidence**

Universal Design

1. **Equitable Use**
2. **Flexibility in Use**
3. **Simple and Intuitive Use**
4. **Perceptible Information**
5. **Tolerance for Error**
6. **Low Physical Effort**
7. **Size and Space for Approach and Use**

Universal Design

- Inclusive assessment population
- Precisely defined constructs
- Accessible, non-biased items
- Amendable to accommodations
- Simple, clear, intuitive instructions and procedures
- Maximum readability and comprehensibility
- Maximum legibility

http://cehd.umn.edu/nceo/TopicAreas/UnivDesign/UnivDesignResources.htm

Assessment for Learning

Assessment should be consistent with our understanding of learning in the subject matter – we need a model of learning to provide a guide for assessment design.

Model of Learning

Effective assessment for learning requires a model of learning the subject matter. For example, research on the development of statistical reasoning or development of specific skills and understanding statistics content will provide the background needed to develop strong assessments.

Model of Learning

A model of learning can describe the learning process, development stages of understanding, knowing, and doing

A model of learning can distinguish novice learners from expert learners; identifying the nature of proficiency and prerequisite skills for progression

Model of Learning

A model of learning allows the test developer to recognize the variety of ways students come to understand the subject matter.

This connects students with the assessment – the assessment reflects the students' experiences.

Model of Learning

A model of learning can provide keys to the kinds of knowledge and skills that are required for achieving content standards or learning objectives.

This allows us to describe the features of tasks that illuminate these aspects of knowledge and skills

Learning Statistics

Ideas from research on teaching & learning:

1. Importance of context
2. Importance of sequencing tasks and knowledge structures
3. Importance of using multiple representations of ideas and concepts

Essential: Purpose

1. Clearly define your purpose
   a. Progress Monitoring (formative assessment)
   b. Objective/Instructional Feedback
   c. Grading (summative assessment)
   d. Placement

Essential: Blueprint

2. Create an assessment blueprint
   a. Content to be covered
   b. Cognitive tasks to be assessed
   c. Format of items
   d. Number of items (given time limits)

Quality MC Items

Content	Knowledge	Comprehension	Application	Total
Central Tendency				25%
Variability				50%
Shape of Distribution				25%
Total	20%	30%	50%	

Essential: Item Quality

3. Design effective items & tasks
   a. Use accepted principles of item writing
   b. Tryout new item types
   c. Review items prior to use – peer review

Writing MC Items

Item Writing Guidelines
- Content Concerns
- Formatting Concerns
- Writing the Stem
- Writing the Choices

Content Concerns

1. Every item should reflect specific content and a single specific mental behavior, as called for in test specifications.
2. Base each item on important content; avoid trivial content.
3. Use novel material to test higher level learning. Paraphrase textbook language or language used during instruction.
4. Keep the content of each item independent from content of other items on the test.
5. Avoid over specific and over general content.
6. Avoid opinion-based items.
7. Avoid trick items.
8. Keep vocabulary simple for the group of students tested.

Formatting Concerns

9. Use of a number of formats is recommended as appropriate given the content and the respondent:
   – the question, completion, and best answer versions of
   – the conventional MC,
   – the alternate choice, true-false (TF), multiple true-false (MTF),
   – matching, and
   – the context-dependent item and item set formats,
   – but AVOID the complex MC (Type K) format.
10. Format the item vertically instead of horizontally.

Style Concerns

11. Edit and proof items.
12. Use correct grammar, punctuation, capitalization, and spelling.
13. Minimize the amount of reading in each item.

Writing the Stem

14. Ensure that the directions in the stem are very clear.
15. Include the central idea in the stem instead of the choices.
16. Avoid window dressing (excessive verbiage).
17. Word the stem positively, avoid negatives such as NOT or EXCEPT. If negative words are used, use the word cautiously and always ensure that the word appears capitalized and boldface.

Writing the Choices

18. Develop as many effective choices as you can, but research suggests three is adequate.
19. Make sure that only one of these choices is the right answer.
20. Vary the location of the right answer according to the number of choices.
21. Place choices in logical or numerical order.
22. Keep choices independent; choices should not be overlapping.

Writing the Choices

23. Keep choices homogeneous in content and grammatical structure.
24. Keep the length of choices about equal.
25. *None-of-the-above* should be used carefully.
26. Avoid *All-of-the-above*.
27. Phrase choices positively; avoid negatives such as NOT.

Writing the Choices

28. Avoid giving clues to the right answer, such as
    a. Specific determiners including always, never, completely, and absolutely.
    b. Clang associations, choices identical to or resembling words in the stem.
    c. Grammatical inconsistencies that cue the test-taker to the correct choice.
    d. Conspicuous correct choice.
    e. Pairs or triplets of options that clue the test-taker to the correct choice.
    f. Blatantly absurd, ridiculous options.

Writing the Choices

29. Make all distractors plausible.
30. Use typical errors of students to write your distractors.
31. Use humor if it is compatible with the teacher and the learning environment.

Item Writing Evidence

- Avoid, or use sparingly, the phrase “none of the above” (57)
- Use as many functional distractors as possible (51)
- Word the stem positively (18)
- State the stem in question form (17)
- Keep the length of options fairly consistent (17)
- Avoid the complex multiple-choice format (13)

Average Effect of Rule Violations

Rule Violation	Difficulty Index	Discrim Index	Reliability Coefficient	Validity Coefficient
Using NOTA	−0.035 (0.005) 51	−0.027* (0.035) 47	−0.001* (0.039) 21	0.073 (0.051) 11
Negative Stems	−0.032 (0.010) 18		−0.168 (0.082) 4	
Open Stem	0.016* (0.009) 17	−0.003* (0.076) 6	0.031* (0.069) 10	0.042* (0.124) 4
Longer correct	0.057* (0.014) 17			−0.265* (0.164) 4
Type-K Format	−0.122* (0.011) 13	−0.146* (0.063) 10	−0.007* (0.083) 4	

Average Effect of Rule Violations

Rule Violation	Difficulty Index	Discrim Index	Reliability Coefficient	Validity Coefficient
Using NOTA				
Negative Stems			**−0.168** (0.082) 4	
Open Stem				
Longer correct	**0.057*** (0.014) 17			**−0.265*** (0.164) 4
Type-K Format	**−0.122*** (0.011) 13	**−0.146*** (0.063) 10		

So what about those distractors?

- Distractors are designed to ***distract*** students, hopefully those with less ability.
- Distractors tend to be fillers, to occupy the other three or four options.
- Distractors are sometimes absurd or even humorous.
- “It’s hard to write 3 or 4 good distractors!”
- Distractors are usually not ***attractors.***

From distractor to attractor

- Attractors ***attract*** students with specific misconceptions or reasoning errors.
- Attractors focus attention on attracting the right students – those with lower ability.
- Attractors require the incorrect options to be plausible, yet not the best answer.
- Attractors are not filler options.

Item Response Attractors

- Attractors contribute to the overall quality of the item and test.
- Attractors play a central role in determining the difficulty of an item.
- Attractors are explicitly designed to inform us about prevailing misconceptions and errors.
- Attractors are explicitly intentional.
- Attractors are explicit.

Improving Diagnostic Information

- Distractors that are written to be plausible should contain common errors or misconceptions
- Distractor analysis provides information regarding the kinds of errors or misconceptions held by students
- No reason, psychometrically, to have the same number of options for every item

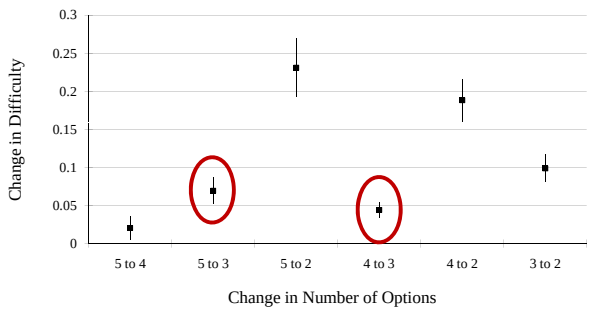
Number of Options

- Less time is needed to prepare 2 plausible distractors than 3 or 4 distractors
- More 3-option items can be administered within the same time limit than 4 or 5-option items, improving content coverage
- Evidence suggests no significant reduction in test item or test score quality by reducing the number of options

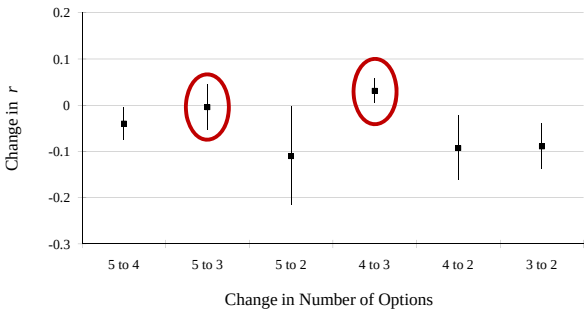
How Many Options? A Meta Analysis

- Subject Area
  - 19 Language Arts
  - 13 Social Science
  - 6 Science
  - 5 Mathematics
  - 3 Mixed Subjects
  - 10 Other (music acoustics, Air Force instruction, entry-level police officer selection, health professional exams)

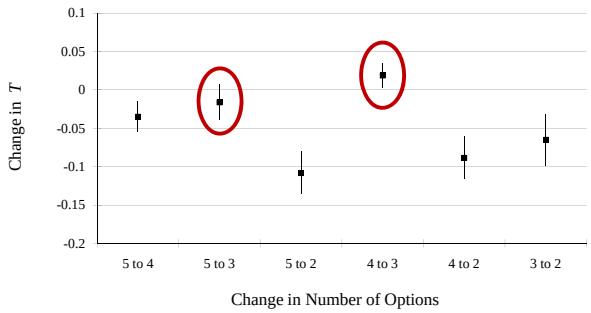
Item Difficulty



Item Discrimination



Test Score Reliability



Issues Related to Access

- Fewer options reduces cognitive load
- More options result in exposing additional aspects of the domain to students – possibly providing clues to other questions
- More options can introduce irrelevant aspects of the domain

Developing CR Items

- Use CR tasks to assess thinking and skills that cannot easily be measured by MC items (worth the cost and effort)
- Assumptions necessary to respond correctly should be related to the content demands of the assessment
- Avoid ambiguous task features – provide full opportunity for students to perform – let them know what is expected

Researchers surveyed 1,000 randomly selected adults in the U.S. A statistically significant, strong positive correlation was found between income level and the number of containers of recycling they typically collect in a week. Please select the best interpretation of this result.

a. We can not conclude whether earning more money causes more recycling among U.S. adults because this type of design does not allow us to infer causation.

b. This sample is too small to draw any conclusions about the relationship between income level and amount of recycling for adults in the U.S.

c. This result indicates that earning more money influences people to recycle more than people who earn less money.

Researchers surveyed 1,000,000 randomly selected adults in the United States. A statistically significant, strong positive correlation was found between income level and the weekly number of containers of recycling collected. Please select the *best interpretation* of this result.

A. The researchers *can* conclude that earning more money influences adults in the United States to recycle more.

B. The researchers *cannot* conclude that earning more money influences adults in the United States to recycle more.

Explain. **(3pts)**

Researchers surveyed 1,000,000 randomly selected adults in the United States. A statistically significant, strong positive correlation was found between income level and the weekly number of containers of recycling collected.

Why is it NOT appropriate for the researchers to conclude that income level causes recycling behavior?

Quality Essay Items

- Restrict the use of essay questions to those learning outcomes that cannot be measured satisfactorily by objective items.
- Construct questions that will call forth the skills specified in the learning standards.
- Phrase the question so that the student's task is clearly indicated.
- Indicate an approximate time limit for each question.
- Avoid the use of optional questions.

Linn and Gronlund (2000)

Essential: Item Review

4. Item Review – Item Analysis
   a. Item Difficulty
      - Proportion that correctly respond
   b. Distractor Functioning
      - Are the distractors being selected
      - Are the distractors "attracting" the right students
      - What misconceptions remain

Assessments for Learning

Formative assessments are specifically designed to support, enhance, and improve learning. Assessments are only formative if they can inform teaching and learning – requiring a feedback loop to students and teachers.

Formative Assessments:

- Provide an organizational framework for content, knowledge, skills – organize content based on the structure of the assessment.
- Confirm "storage" of knowledge by solidifying the connections among different pieces of knowledge.
- Shape study behavior.
- Enhance academic motivation and effort through provision of feedback.

Formative Assessments:

- Enhance the quality and strength of skills by providing unique opportunities to display knowledge and skills.
- Explicitly articulate and communicate learning objectives and achievement targets – typically vaguely defined by teachers.
- Confirm the importance of hard work, time spent studying, and effort.

Formative Assessments:

- Demonstrate the kinds of thinking and processes valued by the instructor.
- Allow students to communicate their thinking about the content and process, convey understanding and misunderstanding.
- Confirm one's own level of understanding and ability to respond on demand.
- Provide opportunities for students to identify their own strengths and weaknesses.

Improving Accessibility

- Test items provide opportunities for students to display construct-relevant knowledge, skills, and abilities.
- From a measurement perspective, item modifications should be done to
  - Provide **access** to the item for all students
  - Improve measurement of the construct
- The hypothesis is: By providing greater **access** to each item, we improve measurement.

Modifications

- changes to a test's content or item format that make a test more accessible for most students
- while continuing to assess grade-level content and skills
- at the same depth of knowledge as unmodified items

Item Modification

- Much of the language surrounding test item modification suggests that the goal is to make items "easier".
- But making items easier doesn't necessarily improve measurement or accessibility.
- Elements of Universal Design provide a model for making appropriate modifications; but can be taken too far to interfere with the construct being measured.

To be Consistent

We are compiling evidence from our own work and the work of others regarding the utility of being direct and explicit in our item writing, relying on good item writing guidelines, elements of Universal Design, concepts from Cognitive Load Theory, and research based on language complexity and accessibility for diverse students.

TAMI provides a systematic guide.

Effects of Item Modification

- Item Difficulty
- Item Discrimination
- IRT Item location and Item Fit
- Distractor Functioning
  - Proportion selecting the distractor
  - Point-biserial correlation (distractor-total *r*)
  - DDF
  - Qualitative Evidence

Consortium for **A**lternate **A**ssessment **V**alidity and **E**xperimental **S**tudies

- Arizona, Hawaii, Idaho, Indiana
- 755 students (WOD, NE, E) in 8th grade
- Develop a common set of items from Discovery Education Assessment reading and mathematics tests
- CBT experimental control of item delivery
- Cognitive Labs

CAAVES Modifications (TAMI)

- Rating accessibility:
  - Passage and stimulus materials
  - Item stem
  - Visuals
  - Answer choices
  - Page format and layout
  - Fairness
- Modification team: 12 teachers, 4 test developers

CAAVES Modifications

- Removal of one option (3-option items)
- Simplification of language (item passage, stem, response options)
- Add graphic support
- Reorganize layout (segment paragraphs, bold key words, add white space)

Overall Effects of Modification

- Effect of Modified Test in Reading
  - Students Without Disabilities: 0.58
  - Students Eligible for Modified Test: 1.24
- Effect of Modified Test in Mathematics
  - Students Without Disabilities: 0.46
  - Students Eligible for Modified Test: 0.81
- Item difficulty was decreased on average
- Item discrimination was increased on average

Focus Group Comments (7th Grade)

- Preparing for State Test:
  - Sleep well, eat good, and stay awake
  - We get snacks during the test
  - To study, I read a lot and do practice questions
  - I don't like to review my notes, it's pretty boring
  - When I get bored I get the temptation for guessing, but I'm not doing that any more
  - We do extra writing in class, teachers take the time to help us practice so we don't forget

Focus Groups cont.

- Reaction to Modified Versions of Items
  - It has a complete question
  - "perplexed"? I have no idea what that means
  - There's always one dumb answer – you don't have to waste your time on that one
  - Just give us the problem straight up, you don't need all of this other information
  - You get the question right at the part of the story where you need to answer it

Focus Groups cont.

- Final Comments:
  - They tell us they are trying to prepare us for college, but they are showing us questions that we never get to in class
  - We shouldn't be completely isolated from all students. If we could use devices to hear music, we wouldn't be distracted.
  - Sometimes the story is so long and you don't have time to read and they only have one or two questions for that story anyway.

Focus Groups cont.

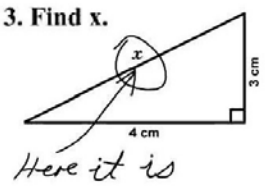
- You should make two kinds of tests. One is the regular old test and then the new and improved test. Then you give it to some kids with disabilities and some without so you can compare. But it has to be random so you know what matters.

CR Item Guidelines (Gitomer, 2007)

For all assessment tasks, regardless of format and response requirements, valid inferences about student understanding require the following:

1. *The student understands what is being asked by the task, including response requirements.*
2. *The scoring system knows how to consistently interpret the student response.*

Essential: Purpose



What is the role of necessary assumptions in order to respond in a way that is consistent with the construct being measured?

CR Guidelines (Gitomer, 2007)

1. **Justify use of CR task formats –** *CR tasks require student thinking that cannot be elicited through fixed-choice formats.*
2. **Inferences should be explicit and construct-relevant –** *Students and assessors should not have to make any inferences about task demands and responses that require implicit assumptions. Task difficulty should be a direct function of knowledge of the construct of interest and avoid manipulation of construct irrelevant factors.*
3. **Distinctions should be clear *across score points* –** *Clear distinctions in the quality of evidence required to satisfy each and every score point are defined.*

CR Guidelines (Gitomer, 2007)

4. **Justifications should be clear within score points –** *Alternative paths to a particular score point ought to be cognitively equivalent.*
5. **Avoid over-specification of anticipated responses –** *Anticipating specific combinations of response features can result in responses that are problematic to assign to rubric score points.*
6. **Empirically verify and modify task through pilot studies**
7. **Beyond the task – construct measurement should be consistent across tasks –** *Expectations for the same cognitive aspects should be consistent across tasks within an assessment, as appropriate.*

NAEP

- National Assessment of Educational Progress
- Examples From the Science Assessment
- Grades 8 and 12
- 2000-2005