Michael C. Rodriguez

Measurement Essentials that Support ***Assessment for Learning***

EPSY 5221: Principles of Educational & Psychological Measurement

Measurement

Enduring Themes:

What to measure

How to measure

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

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 Models

Ideas from research on 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 uses)
   b. Objective/Instructional Feedback
   c. Grading (summative uses)
   d. Placement

Essential: Blueprint

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

Quality 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 and expert review

Mr. Wilson and 3 friends dined at a popular restaurant. The bill was $77 and they left a $15 tip. Approximately what percentage of the total bill did they leave as a tip?

A. 10%
B. 13%
C. 15%
D. 20%
E. 25%

Embretson (2007)

A total of 80 players were in a football league. There were 10 players on each team. Which number sentence is in the same fact family as 80 ÷ 10 = 8?

A. 8 × ? = 80

B. 8 × 10 = ?

C. ? × 80 = 10

D. 8 × 80 = ?

Embretson (2007)

Writing MC Items

- Questions should require students to consider novel contexts
- Use reference materials (graphical displays) that are authentic
- Options should be plausible – common errors or misconceptions
- Use only the number of options you need or can develop (3 is sufficient)

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?

The Power of Distractors

- Item difficulty is often determined by the distractors
- Distractor proximity and plausibility determine the difficulty of items

Who was the 7th president of the United States?

A. Andrew Jackson
B. John Quincy Adams
C. Abraham Lincoln
D. James Madison

Who was the 7th president of the United States?

A. Andrew Jackson

B. Davy Crockett

C. George H. W. Bush

D. Elvis Presley

A high school graduation test is considered a criterion-referenced test because

A. it was built with a test blueprint.
B. most students will pass a graduation test.
C. it is based on a clearly defined domain.

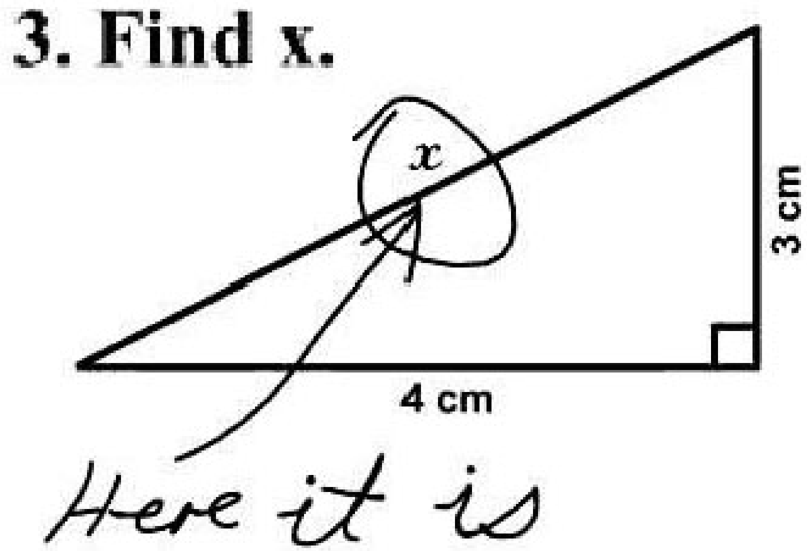
A high school graduation test is considered a criterion-referenced test because

A. it was really difficult.

B. teachers spent a month on test preparation.

C. it is based on a clearly defined domain.

Be Explicit



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

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

Essential: Item Review

4. Item Review – Item Analysis

   c. Item Discrimination

   - Item-Total correlation (point-biserial)
   - Heuristic: high ability students will tell us which option is correct

Essential: Score Quality

5. Assessing score reliability
   a. Test-retest consistency
   b. Internal consistency
   c. Inter-rater agreement
   d. Person×Task variance (G-Theory)

Assessments for Learning

Assessments can be designed to support, enhance, and improve learning.

Assessments are only formative if they inform (change) teaching and learning – requiring a feedback loop to students and teachers.

Formative Uses of Assessment:

- 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 Uses of Assessment:

- 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 Uses of Assessment:

- Demonstrate the kinds of thinking and processes valued by the instructor – valued by the field.
- 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.

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 being claimed?”

- **Develop Assessments to gather Evidence**

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

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.

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	

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

Educational Measurement
ISSUES AND PRACTICE

NCME national council on measurement in education

Three Options Are Optimal for Multiple-Choice Items: A Meta-Analysis of 80 Years of Research

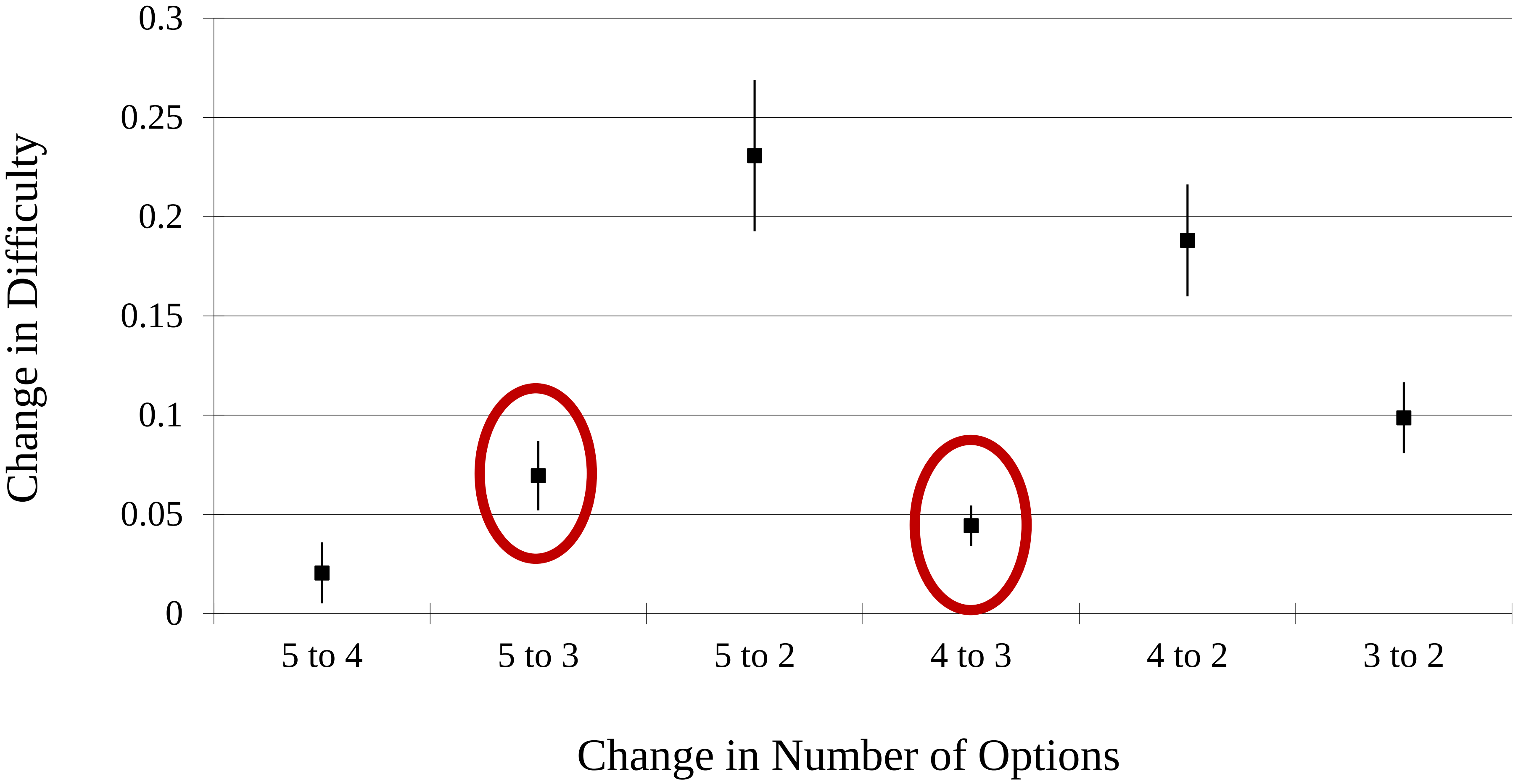
Michael C. Rodriguez, *University of Minnesota*

2005. Volume *24*(2), 3-13.

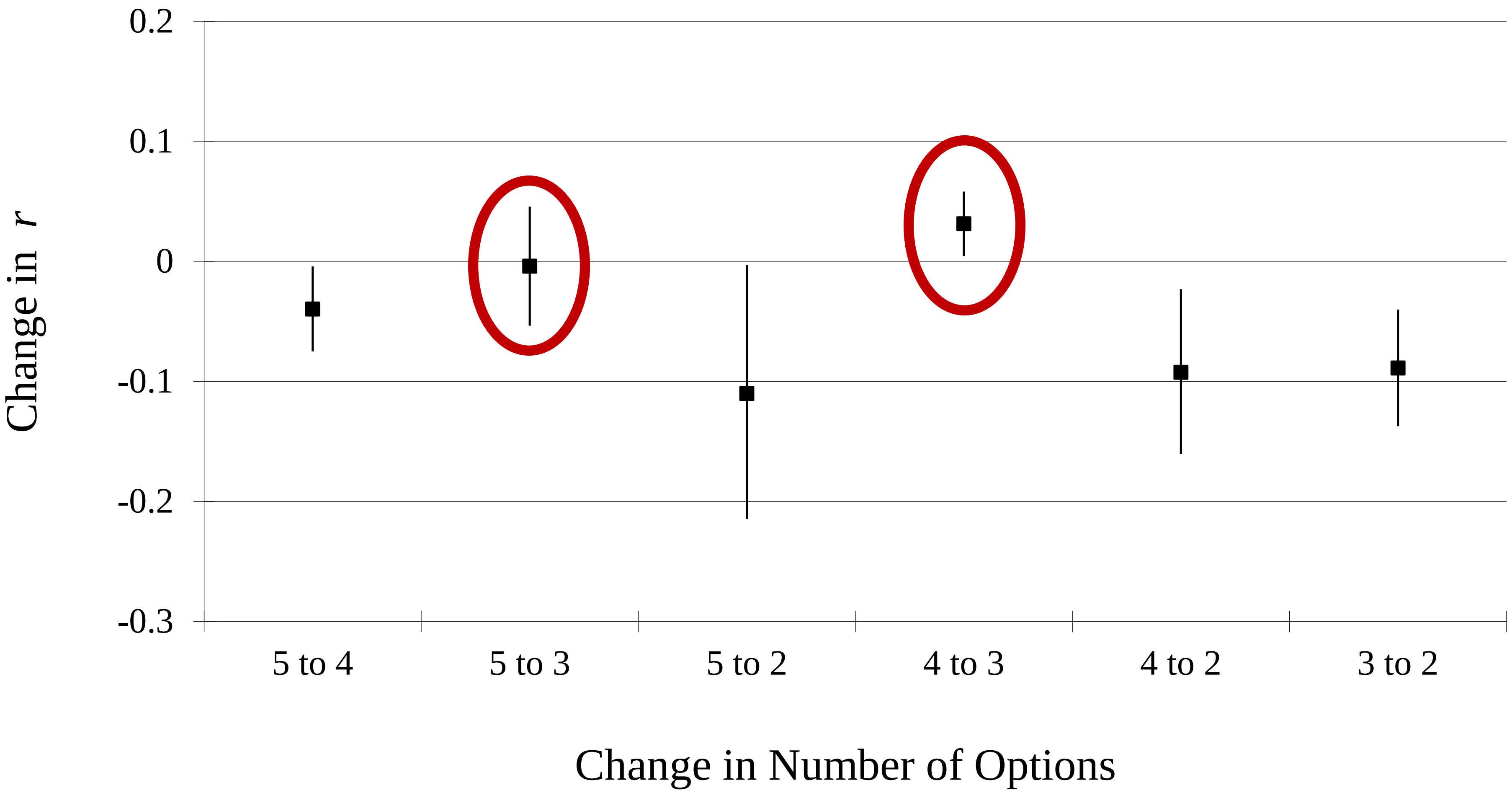
How Many Options?
A Meta Analysis

- 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)

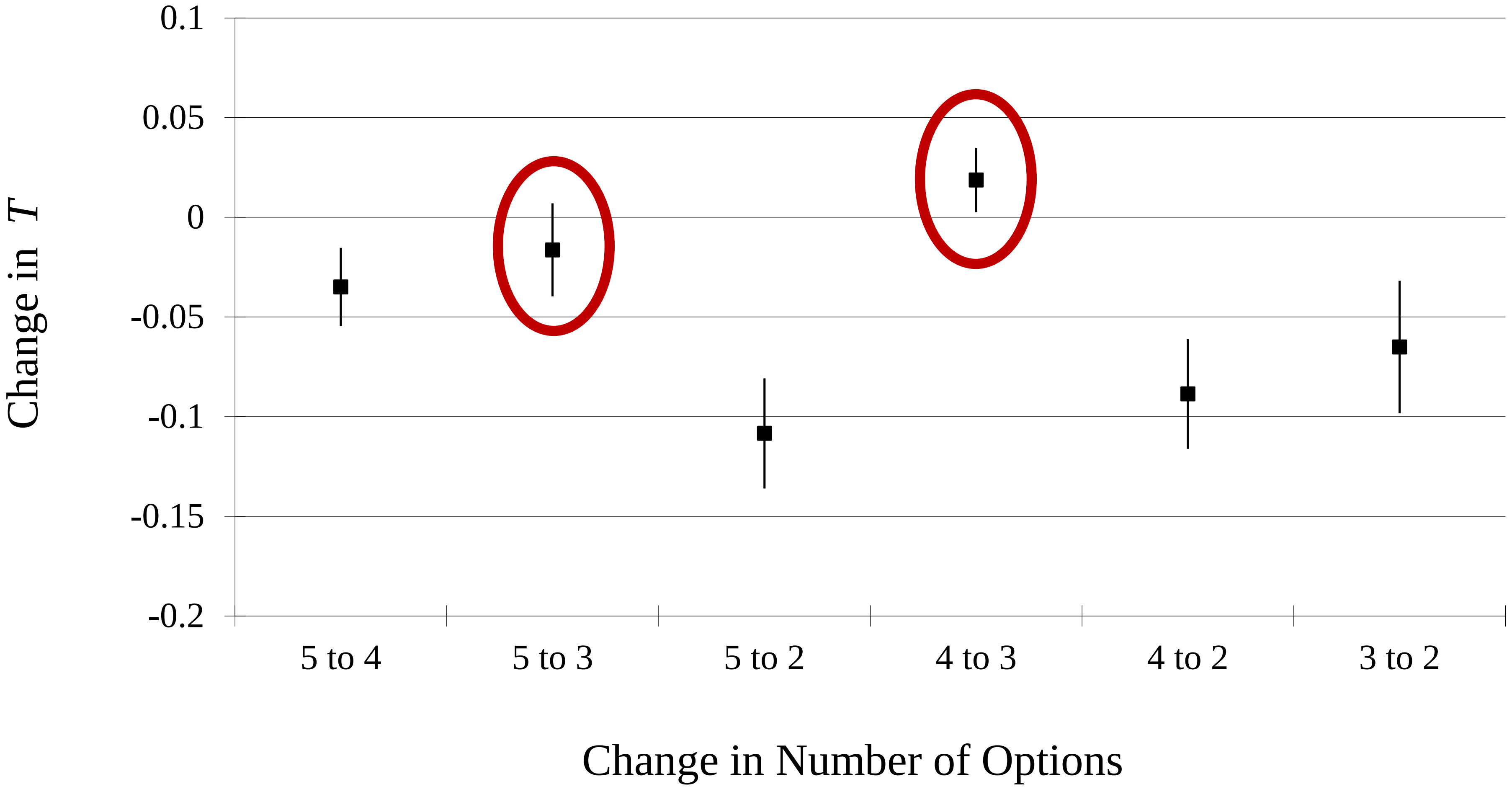
Item Difficulty



Item Discrimination



Test Score Reliability



Implications for Accessibility

- 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

References

Embretson, S.E. (2007). Construct validity: A universal validity system or just another test evaluation procedure? *Educational Research, 38*, 449-455.

Rodriguez, M.C. (2005). Three options are optimal for multiple-choice items: A meta-analysis of 80 years of research. *Educational Measurement: Issues and Practice, 24*, 3-13.

Haladyna, T.M., Downing, S.M., & Rodriguez, M.C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education, 15*, 309-334.

Linn, R.L. & Gronlund, N.E. (2000). *Measurement and Assessment in Teaching*. Upper Saddle River, NJ: Prentice-Hall.

Mager’s Objectives

Instructional objectives specify what a student must do to demonstrate learning:

- Process, including who, what, and when.
- Product, indicating what a student must do or know as a result of instruction.

Mager’s Objectives

- When clearly defined objectives are lacking, there is no sound basis for the selection or designing of instructional materials, content, or methods. If you don't know where you are going, it is difficult to select a suitable means for getting there.

http://www2.gsu.edu/~mstmbs/CrsTools/Magerobj.html

Mager’s Objectives

- Tests or examinations are the mileposts along the road of learning and are supposed to tell instructors AND students whether they have been successful in achieving the course objectives. But unless objectives are stated clearly and are fixed in the minds of both parties, tests are at best misleading; at worst, they are irrelevant, unfair, or uninformative.
- Test items designed to measure whether important instructional outcomes have been accomplished can be selected or created intelligently only when those instructional outcomes have been made explicit.

http://www2.gsu.edu/~mstmbs/CrsTools/Magerobj.html

Mager’s Objectives

Mager has three elements of instructional objectives:

- Behavior or overt activity to be performed by student (the verb must be observable)
- Specification of conditions under which the behavior is observed
- A minimum level of acceptable performance

Mager's Objectives

EXERCISE:

Identify the missing component:

A = Behavior

B = Conditions

C = Criteria

Mager’s Objectives

Poorly written objectives:

- The students will appreciate poetry.
- After the course, students will be better readers.
- Students will write a complete sentence.
- Students will become better neighbors.
- Students will understand and value the need for recycling.

Mager’s Objectives

- Given ten rocks, the student will label them as igneous, metamorphic, or sedimentary rocks.
- The student will write a sonnet that follows the proper form.
- Given three 7-word sentences, the student will correctly identify the parts of speech for 18 of the words.

Mager’s Objectives

- The student will locate 12 major bones on the diagram of the skeleton.
- The student will identify the faults in a poorly written objective.
- The student will recognize ten instances of positive reinforcement.
- Given five sentences, the student will correctly classify four of them.

Mager’s Tips on Objectives

http://www2.gsu.edu/~mstmbs/CrsTools/
Magerobj.html

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

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.*

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)

JEM JOURNAL of EDUCATIONAL MEASUREMENT

NCME national council on measurement in education

Journal of Educational Measurement
Summer 2003, Vol. 40, No. 2, pp. 163–184

Construct Equivalence of Multiple-Choice and Constructed-Response Items: A Random Effects Synthesis of Correlations

Michael C. Rodriguez
University of Minnesota

Publication Date of Empirical Studies Located in This Search

Number of Studies	1920s	1930s	1940s	1950s	1960s	1970s	1980s	1990s
Included in Synthesis	2	3	0	2	5	4	6	7
Not included	0	0	0	0	3	2	9	18

Mean Correlations given Test Design

Summary of Moderator Analysis for Weighted Average Corrected Correlations.

n	Design	Mean *r*	Mean *r*	Mean *r*	Mean *r*
21	Stem-Equivalent	.949	.949	.949	.949
12	Content-Equivalent	.916	.916	.861	.890
18	Not Content-Equivalent	.839	.821		
15	Essay-Type Items	.810			.810

Stem-Equivalent: MC and CR items have identical stems.
Content-Equivalent: MC and CR items written to measure same content.
Not Cont-Eq: MC and CR items written to measure different aspects of the content.
Essay-Type: MC items were compared to essay task scored with a rubric.