Workshop I: Assessment Instrument Design

Clarissa Dirks
The Evergreen State College
Olympia, Washington

Discipline-Based Education Research (DBER)

- Workshop I: Moving from Teaching to Research about Teaching and Learning
- Workshop II: Conducting Discipline-Based Education Research
- **Workshop III: Instrument Design**

Why Design an Assessment Instrument/Tool?

- **No other assessment instrument exists for your needs**
- **Your assessment instrument will be used in a way that is different from similar tools**
- **It will be used by you and many others (or not)**
- **You have copious spare time**

The Workshop of A Million Questions To Which You Have The Answers

If You Must Design an Assessment Instrument Then there are Some Things You Need to Consider

Have you determined what you REALLY want to assess? (What are the very specific learning outcomes, attitudes etc… that you will be assessing?)

Who will use the instrument? (R1 faculty, CC faculty, etc…)

Who is the target audience? (Large classrooms, undergraduates, seniors only, etc…)

Under what conditions will the instrument be used? (homework, in-class, computer-based, etc…)

Group Work (20 minutes)
Keeping your research problem in mind, work in teams of two to answer these questions:

Have you determined what you REALLY want to assess? (What are the very specific learning outcomes, attitudes etc… that you will be assessing?)

Who will use the instrument? (R1 faculty, CC faculty, etc…)

Who is the target audience? (Large classrooms, undergraduates, seniors only, etc…)

Under what conditions will the instrument be used? (homework, in-class, computer-based, etc…)

In What Structure or Format Will Your Test Be?

- Multiple-choice
- Free-response
- Likert scale
- Multiple True/False
- Question clusters (factor analysis)

Group Work (10 minutes)
Discuss the format you would most likely select and why that format is best.

Data Collection: What Instrument Will You Use?

Whatever you use, it should be valid and reliable.

Validated instruments measure what they are supposed to measure, and reliable instruments consistently distinguish between individuals with disparate abilities.

There are many forms of validity and reliability!

Alias, M. Assessment of Learning Outcomes: Validity and Reliability of Classroom Tests. World Transactions on Engineering and Technology Education. Vol 4 No. 2, 2005

Some Important Terminology

Item:
An individual question or task to which a student will respond.

Item Difficulty:
The percentage of students answering the item correctly (a confusing but unfortunately established convention: note that the easiest items have the highest item difficulty).

Item Discrimination:
The item's ability to discriminate between students of different academic ability—those that have mastered the material and those that haven't; the students with mastery of the content are more likely to answer an item correctly.

Validity: Addresses whether the test measures what it is supposed to measure.

Construct Validity: The *item* measures what it is intended to measure (requires affirmation by experts).

Content Validity: The *test* presents the key concepts and misconceptions for a given domain as affirmed by experts. Content validity increases as the number of relevant items on the test increases.

Face Validity: The test has the appearance of measuring what it intends to measure, thereby motivating students to do their best. The assumption is that students performed to the best of their abilities on the test.

Work in small groups to define the following:

Validity:

Construct Validity:

Content Validity:

Face Validity:

Reliability: Addresses the consistency of a set of measures and is based on whether the test results are repeatable and not subject to random error.

Alternative Form Reliability: A student will have relatively the same result when they take alternative forms (isomorphic forms) of the same test at two different times because the different items assess the same content or skills at the same cognitive level.

Internal Consistency: Measures whether pair-wise items on a test that should measure the same content, or latent variable, actually do so.

Inter- or Intra-Rater Reliability: The consistency of scoring free-response items by different graders or by the same grader, respectively.

Test-Re-Test Reliability: A student or group of students will score similarly if they take the same test at two different times without intervention.

Work in small groups to define the following:

Reliability:

Alternative Form Reliability:

Internal Consistency:

Inter- or Intra-Rater Reliability:

Test-Re-Test Reliability:

Two main statistical frameworks that test designers have used: Classical Test Theory (CTT) and Item Response Theory (IRT).

CLASSICAL TEST THEORY

Classical test theory relies on test-score models, where the test scores for a given test group are analyzed and items are modified on the basis of the group results. Therefore, there is great emphasis on finding and using the most suitable group for test development in classical test theory.

ITEM RESPONSE THEORY

In contrast, IRT focuses on the analysis of each item in relation to an examinee's ability and is considered to be test-independent. Examinees come to the test with ability levels observable only through statistical analysis of their performance on each test item

Science Process And Reasoning Skills Test
SPARST

C. Dirks and M.P. Wenderoth

SPARST ADVISORY BOARD

Michelle Withers- West Virginia University
Mark Hens- Univ. North Carolina- Greensboro
Michael Henna-Rensselaer Polytechnic Institute
Kristina Garza- Univ. Texas-El Paso

Ann Murkowski- North Seattle Community College
Jenny McFarland- Edmonds Community College
April Fong- Portland Community College
Lianne Etchberger- Utah State University

What are the benefits of using SPARST?

- **It is a biology content-independent but context-dependent test**
- **The test is designed for biology majors**
- **It is a free, online, multiple-choice test**
- **The test discriminates between biology majors across academic years**
- **The test is modular and testing skills in: experimental design, data analysis, graphing and science communication**

SPARST Development Steps

Preliminary Work to Determine Skills to Be Tested

- In class diagnostics and student interviews
- Evaluation of biology exams
- National faculty survey
- Literature review of student misconceptions
- Analysis of preexisting skills assessments

Question Development

Questions were designed to be context dependent and content independent

Student Readability

Students took the tests (each separately), discussed with a recorder why they selected the option they did, and offered suggestions for improving clarity. We made minor edits.

Advisory Board Review
For Construct and Item Validity

Eight faculty members from a variety of institutions evaluated the graphing, data analysis and experimental design tests and made suggestions for changes. They also identified a few skills we should assess. Faculty brainstormed about the items that should be included in the Science Communication test.

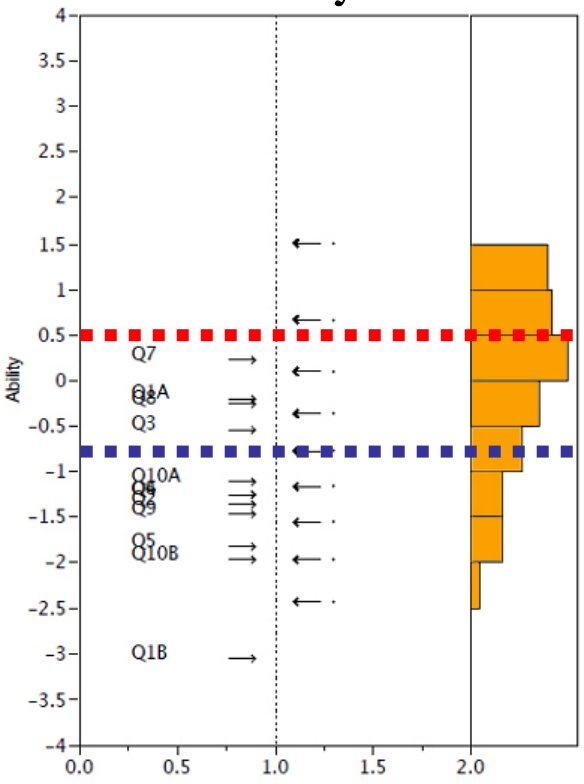
Student Readability
Student Interviews and Revisions

Fall 2012 Pilot
four institutions; 100 and 400 level courses; N = 1094

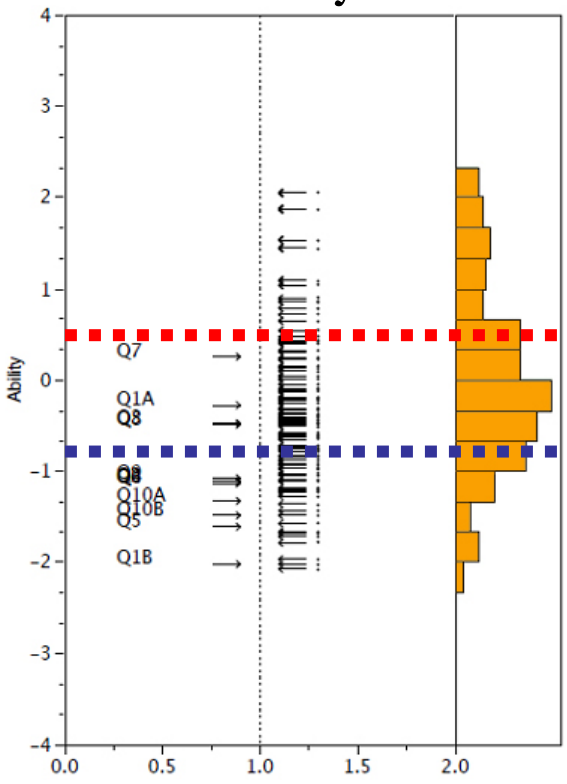
Data Analysis
- Item Response Theory modeling
- Analysis of Wright maps for question stability

Revisions of Items
Rewrite unstable questions (those lacking strong discrimination)

Data Analysis 1-PL



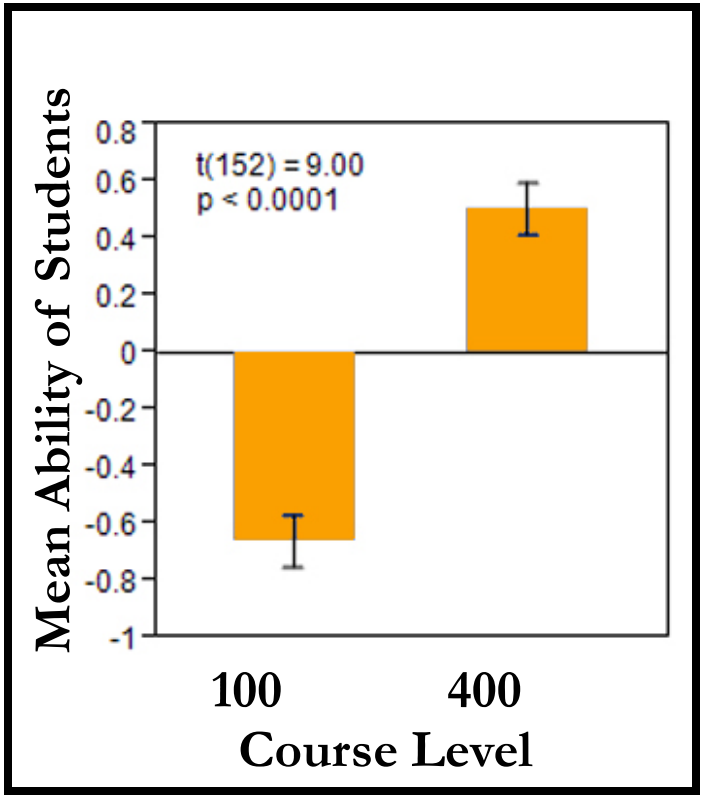
Data Analysis 2-PL



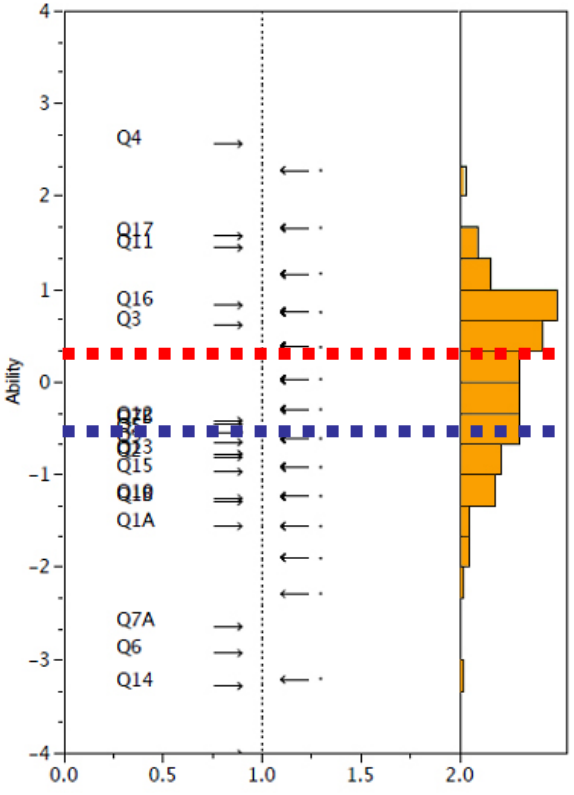
What Does The Data Look Like?

400-level student mean

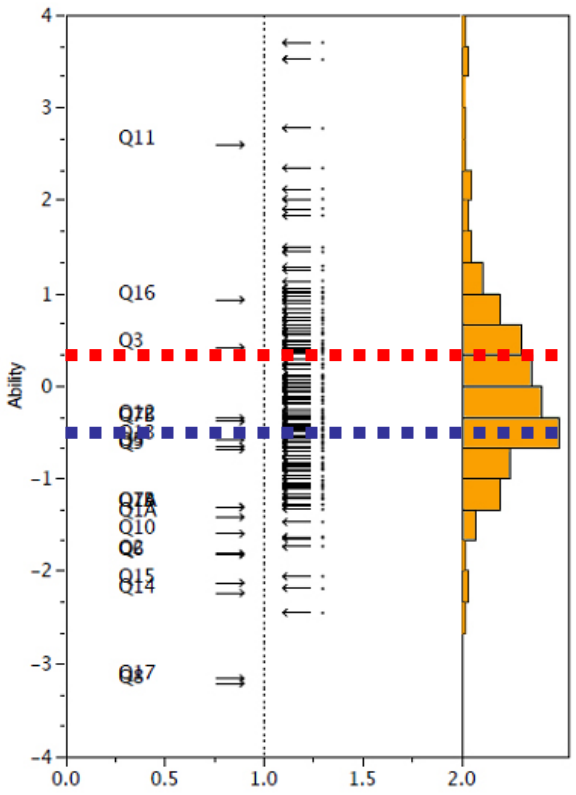
100-level student mean



Graphing 1-PL



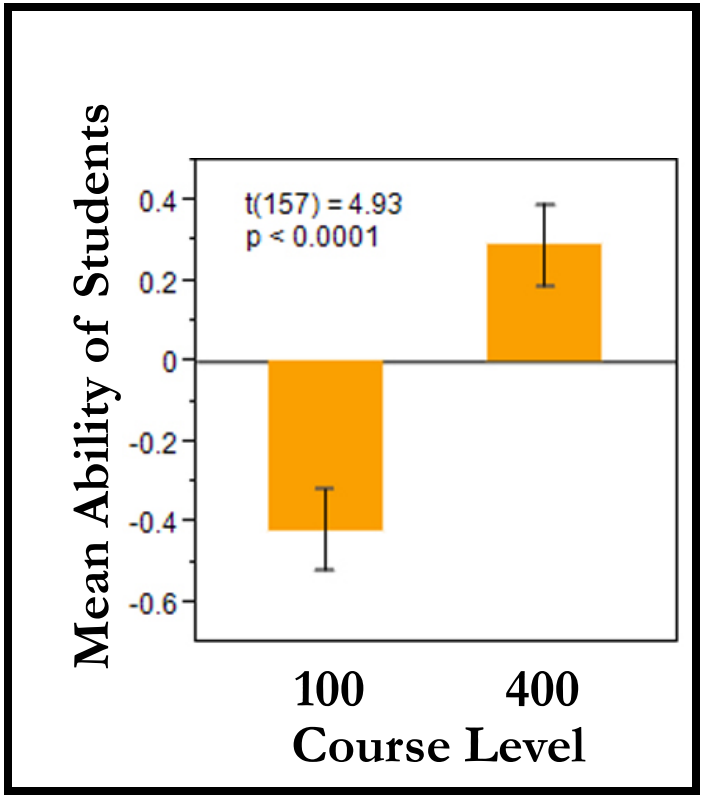
Graphing 2-PL



What Does The Data Look Like?

400-level student mean

100-level student mean



So Do You Really Want to Design an Instrument?

Work in small groups to refine your DBER plan

a) Research idea
b) Research design
c) Collection of data
d) Analysis of data

You Were a Captivating Audience – THANK YOU!!!