

Scaling and Equating

Wendy Yen
Educational Testing Service

New York State
Technical Conference
October 21, 2002



Scaling: What is it?

- Statistical procedure for analyzing test performance (and sometimes item performance)
- When procedure is followed, it produces numbers related to test scores (and sometimes items) that have a particular meaning
- The meaning associated with the scale scores depends on the scaling procedure

Equating: What is it?

- Statistical procedure for measuring and controlling for variations in the difficulty (and other statistical characteristics) of different tests
- Scores from equated tests have comparable meaning

Examples of Scales

- Population-referenced scales for tests
 - Percentiles, grade equivalents, Thurstone scaling, IQ scales
- Non-population-referenced scales for items and tests
 - Number-correct scores
 - Item response theory (IRT)

Why use IRT scaling?

- IRT scale scores and item parameters have a lot more meaning than number-correct scores and other raw scores
- IRT scale scores don't force scale score distributions to have a particular shape
- IRT values are useful for
 - Test construction
 - Test equating
 - Test score interpretation

IRT: Unique Characteristics

- Focuses on items, not just intact tests
- Describes item performance at each level of student ability or achievement ("scale score")
- Is based on a statistical model
 - Describes item performance succinctly
 - Makes it possible to generalize from one testing situation to another

IRT: Useful for Test Construction

- Places item difficulty and student performance on the same scale
 - Works for both multiple-choice & constructed-response items
- Tells how much information each item, or item score level, contributes to test
- Describes differential item functioning for groups of students (e.g., boys and girls)
- Shows impact of using different items in a test
 - Helps test developer create parallel test forms
 - Helps test developer pick items targeted to particular student achievement levels

IRT: Useful for Test Equating

- Places tests with different items on same scale
- Adjusts for difference in test difficulty
 - Students' scale scores (on the average) don't depend on which test form they take
 - Students' scale scores from different test forms are comparable

IRT: Useful for Test Scoring and Interpretation

- Describes the amount of measurement error in each score
- Can provide statistically optimal item weights that produce the most accurate scores
- By placing items and student performance on same scale, IRT
 - Facilitates standard setting
 - Can be used in criterion-referenced score interpretation

IRT Model Fit

- Models make simplifying assumptions
- Some models make stronger assumptions than others
- Accuracy of assumptions must be evaluated before relying on models

Examples

- How item difficulties are placed on the same scale as student performance for
 - Multiple-choice items (Fig. 1-3)
 - Constructed-response items (Fig. 4)
 - Both item types together (Fig. 5)

Examples for a 5-item Test

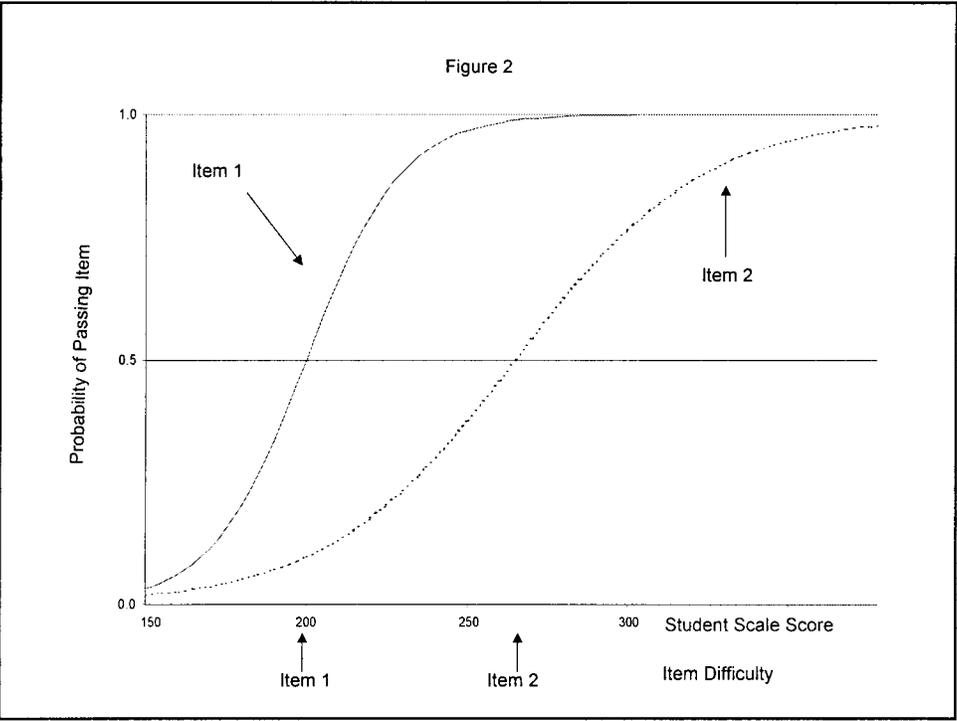
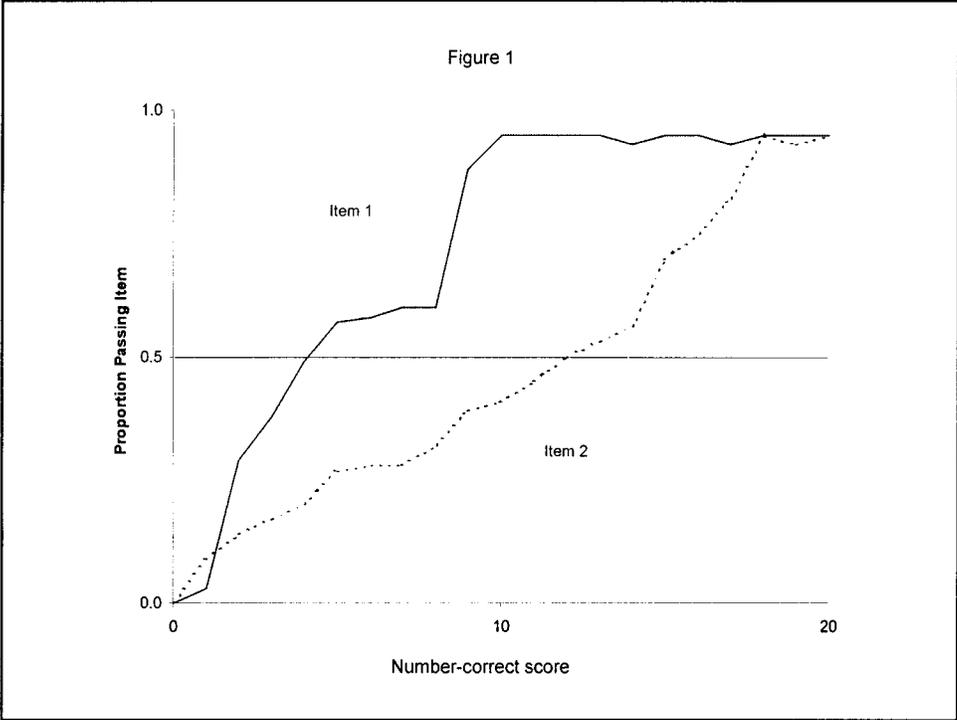
- How students' raw scores (e.g., number-correct scores) are turned into scale scores (Fig. 6 & Table 1)
- How measurement error can be estimated for every scale score (Fig. 7)
- How items located along a scale can facilitate standard setting (Fig. 8)

Examples (cont.)

- How equating is conducted (Fig. 9)
 - Use of anchor items
- How the equating adjusts for differences in item difficulty in different test forms (Fig. 10 and Table 2)
 - The raw score needed to obtain a scale score is appropriately adjusted for differences in item difficulty across forms
 - The scale score cut-point for a standard has the same meaning across forms

Examples (cont.)

- How IRT scaling and equating do not constrain the distribution of student scores to have any particular level or shape
 - Tables can be created that show how each raw score is converted to a scale score (Table 2)
 - These tables can be & commonly are created in advance of operational testing
 - Students' raw scores (actual performance) determine the distribution of scale scores and percents of students reaching each performance level or standard (Fig. 11-13)



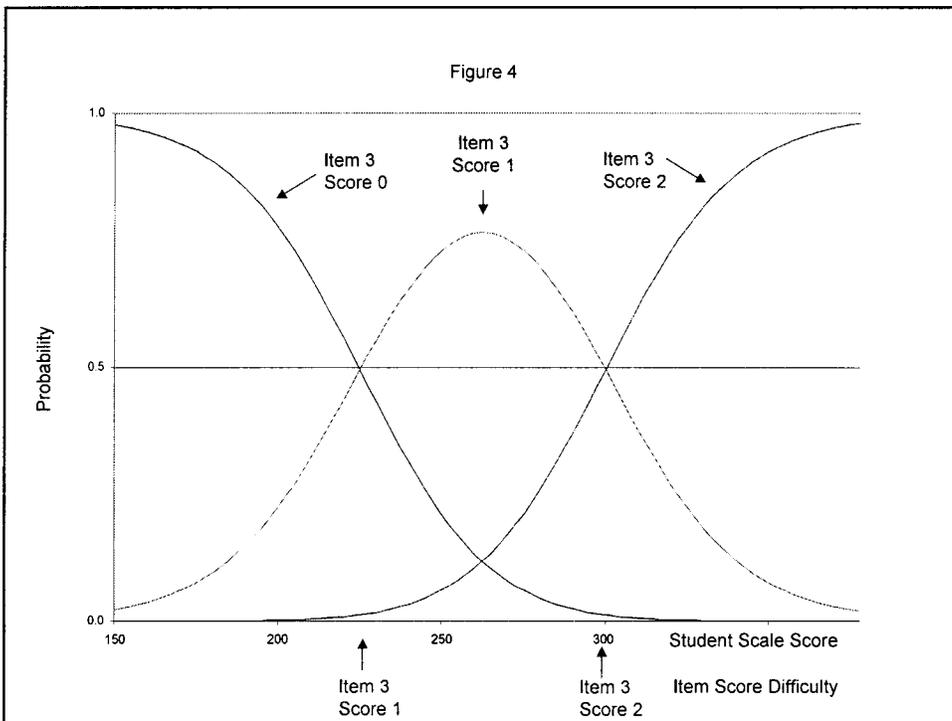
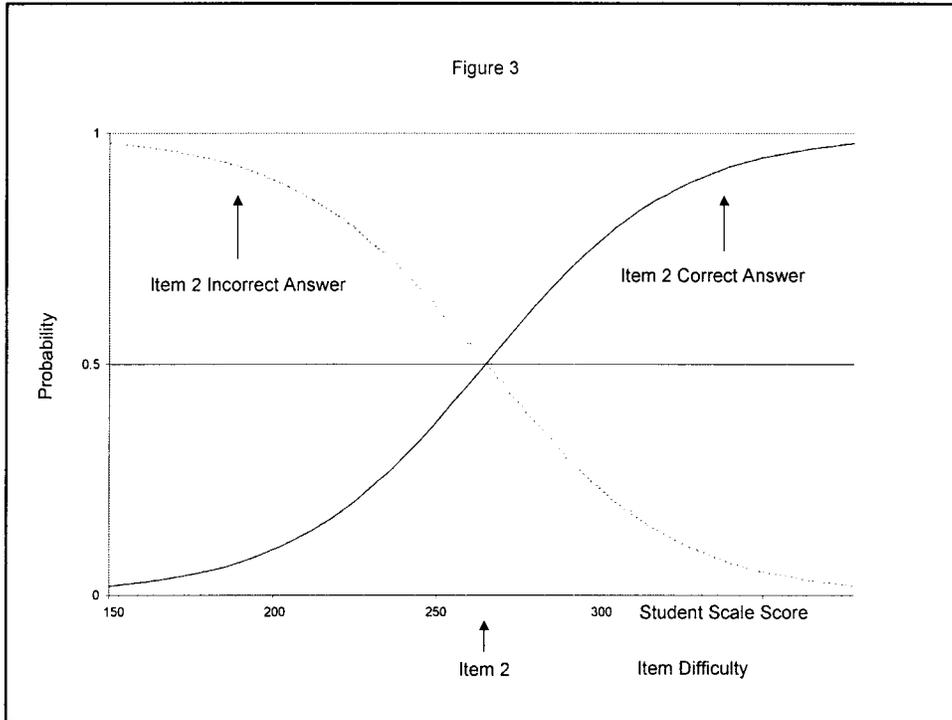


Figure 5

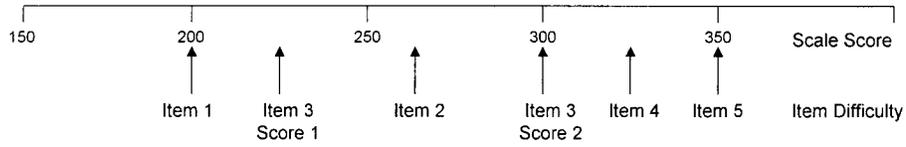


Figure 6

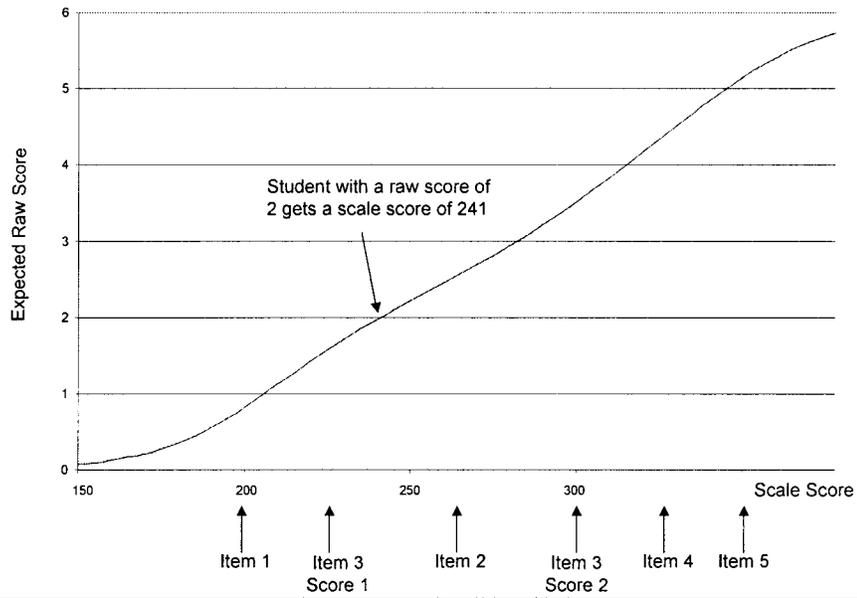


Table 1
Example of Raw Score to Scale Score
Conversion

Raw Score	Scale Score
0	150
1	206
2	241
3	282
4	315
5	345
6	400

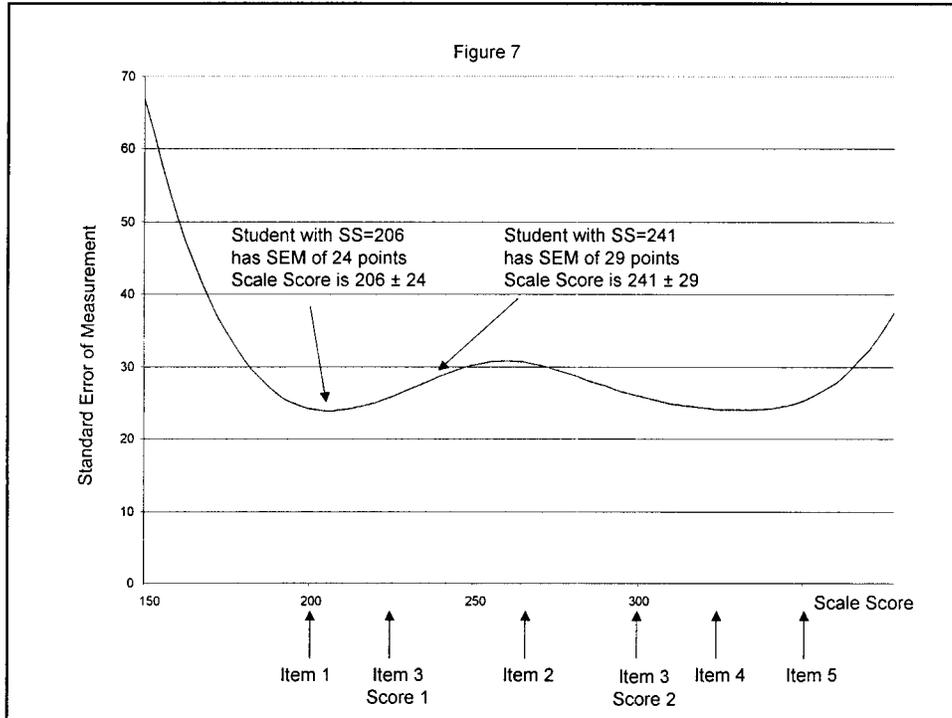


Figure 8

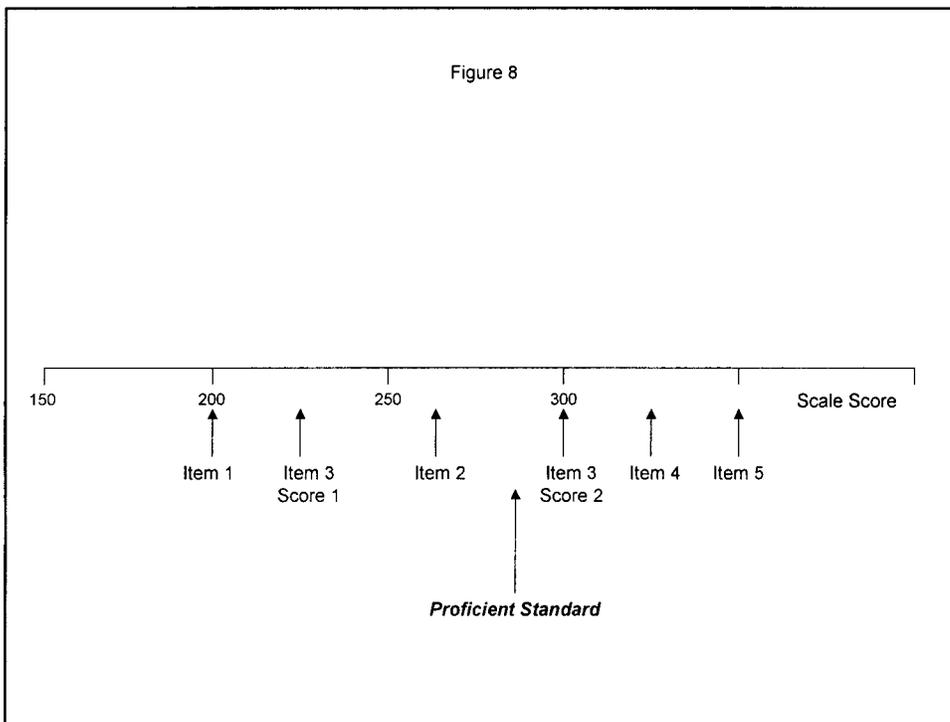
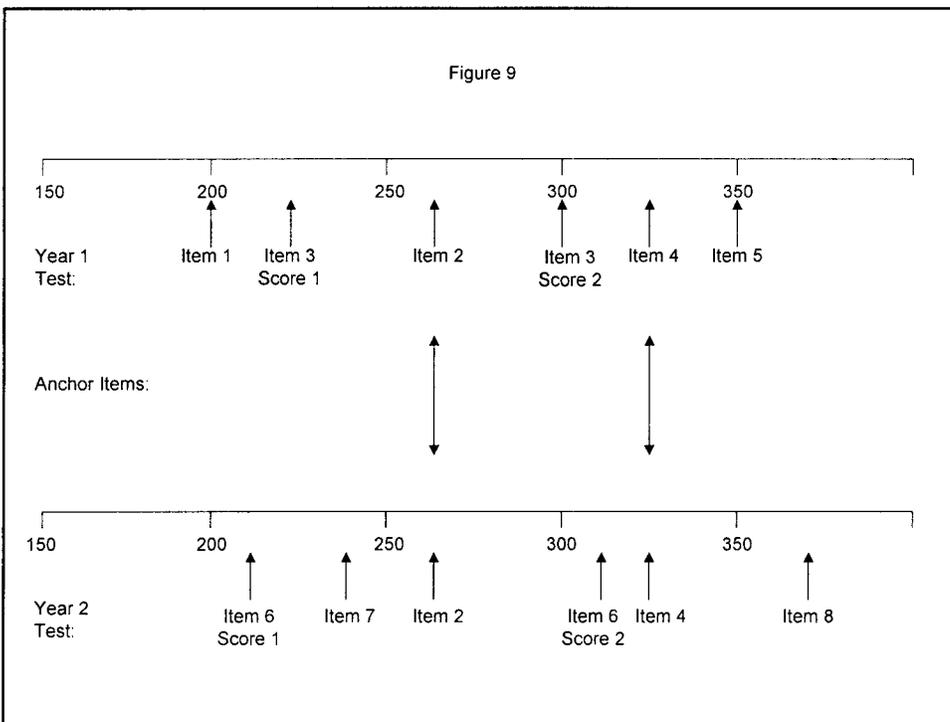


Figure 9



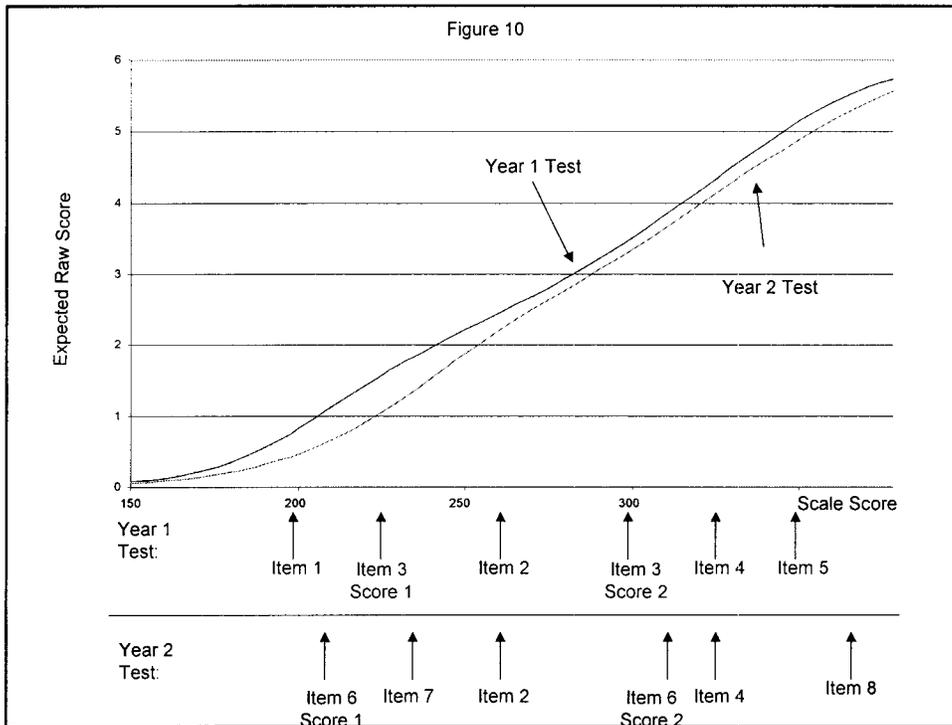
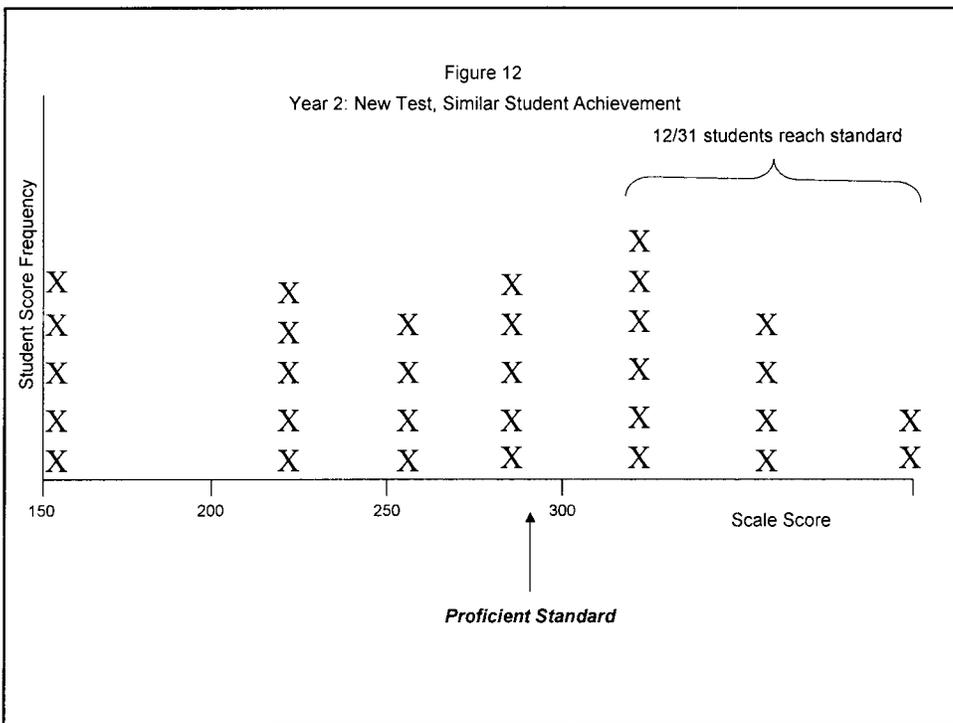
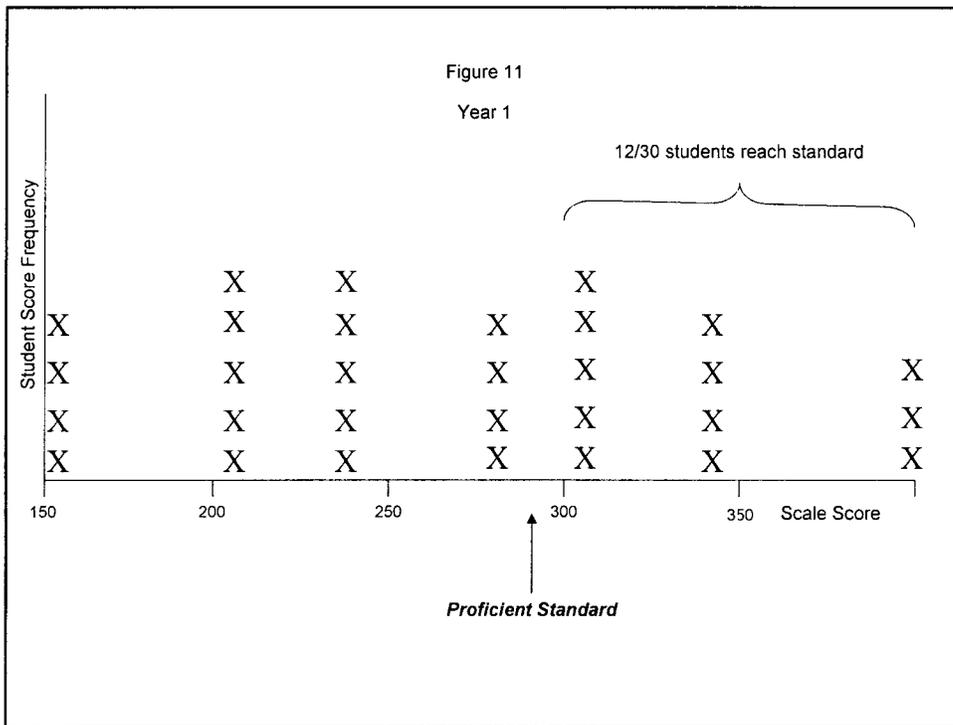
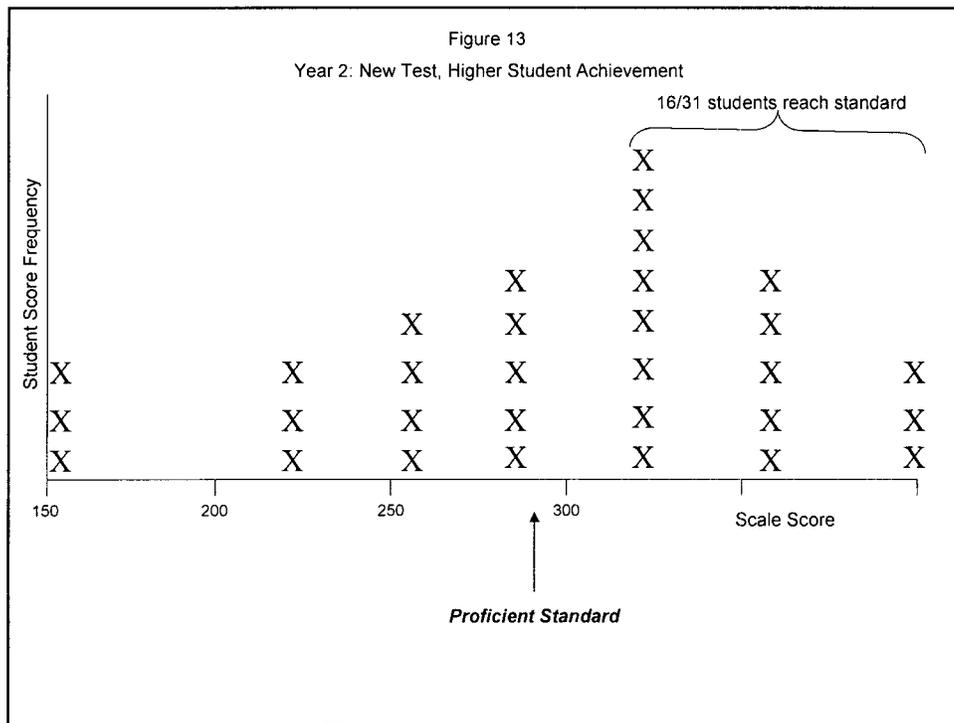


Table 2
Example of Raw Score to Scale Score
Conversions for Two Tests

Raw Score	Scale Score	
	Year 1 Test	Year 2 Test
0	150	150
1	206	223
2	241	254
3	282	288
4	315	321
5	345	354
6	400	400





Further Reading

- “Item Response Theory” in Encyclopedia of Educational Research