

Test Scaling and Value-Added Measurement

Dale Ballou

Vanderbilt University

April, 2008

- VA assessment requires that student achievement be measured on an interval scale: 1 unit of achievement represents the same amount of learning at all points on the scale.
- Scales that do not have this property:
 - Number right
 - Percentile ranks
 - NCE (normal curve equivalents)
 - IRT “true scores”
- Scales that *may* have this property
 - IRT ability trait (“scale score”)

Item Response Theory Models

- One-parameter logistic model

$$P_{ij} = [1 + \exp(-D(\theta_i - \delta_j))]^{-1},$$

P_{ij} is the probability examinee i answers item j correctly

θ_i is examinee i ability

δ_j is item j difficulty

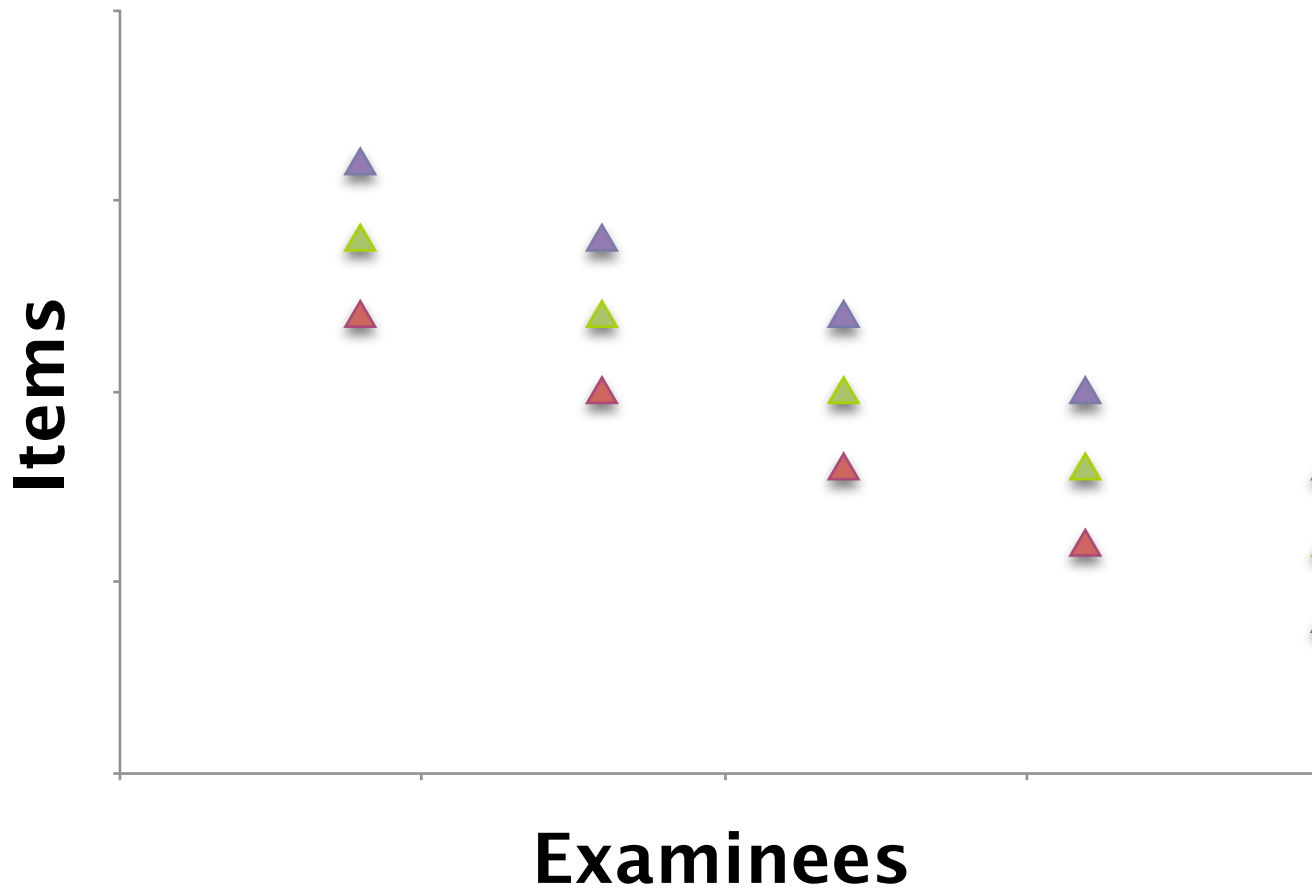
Two- and Three-Parameter Logistic IRT Models

- $P_{ij} = [1 + \exp(-\alpha_j(\theta_i - \delta_j))]^{-1}$
- $P_{ij} = c_j + (1 - c_j)[1 + \exp(-\alpha_j(\theta_i - \delta_j))]^{-1}$

α_j is an item discrimination parameter

c_j is a guessing parameter

IRT Isoprobability Contours (1-parameter model)



Linear, parallel isoproprobability curves are the basis for the claim that ability is measured on an interval scale.

- The increase in difficulty from item P to item Q offsets the increase in ability from examinee A to B, from B to C, and from C to D.
- In this respect, $AB = BC = CD$, etc.
- Moreover, the same relations hold for any pair of items.

Do achievement test data conform to this model?

- P_{ij} and isoprobability contours aren't given. Data are typically binary responses.
- Testable hypotheses can be derived from this structure, but power is low.

- The model doesn't fit the data when guessing affects P_{ij} .
- Or when difficulty and ability are multidimensional.
- “Data” are selected to conform to the model
→ ability may be too narrowly defined.

Implications:

- It seems unwise to take claims that ability is measured on an interval scale at face value.
- We should look at the scales.

CTB/McGraw-Hill CTBS Math (1981)

Grade	Mean gain from previous grade	Standard deviation
2	186	77
3	67	44
4	33	35
5	23	24
6	14	20
7	6	23

CTB/McGraw-Hill Terra Nova, Mean Gain From Previous Grade (Mississippi, 2001)

Grade	Reading	Language Arts	Math
3	23.8	30.2	47.3
4	21.5	20.4	24.7
5	15.5	17.8	19.0
6	10.5	6.9	21.7
7	9.2	9.9	10.7
8	12.4	10.9	17.0

Northwest Evaluation Association, Fall, 2005, Reading

Grade	Mean gain from previous grade	Standard deviation
3	14.7	15.6
4	9.5	15.0
5	6.9	14.6
6	4.8	14.8
7	4.0	14.8
8	3.5	14.8

Northwest Evaluation Association, Fall, 2005, Math

Grade	Mean gain from previous grade	Standard deviation
3	14.0	12.1
4	10.9	12.8
5	8.5	13.9
6	6.4	15.0
7	5.8	15.0
8	4.8	16.8

Appearance of Scale Compression

- Declining between-grade gains
- Constant or declining variance of scores

Why?

- In IRT, the same increase in ability is required to raise the probability of a correct answer from .2 to .9, regardless of the difficulty of the test item. Do we believe this?

To raise the probability of a correct response from $\frac{2}{7}$ to 1, who must learn the most math?

Student A

What makes us think of a circle?

- A. Block
- B. Pen
- C. Door
- D. A football field
- E. Bicycle wheel

Student B

Using the Pythagorean Theorem, $a^2 + b^2 = c^2$, when $a = 9$ and $b = 12$, then $c = ?$

- A. 8
- B. 21
- C. 15
- D. $\sqrt{21}$
- E. 225

Responses

- Conference Participants
 - A: 11
 - B: 26
 - Equal: 7
 - Indeterminate: 30
- Faculty and Graduate Students, Peabody College
 - A: 13
 - B: 37
 - Equal: 15
 - Indeterminate: 33

Implications

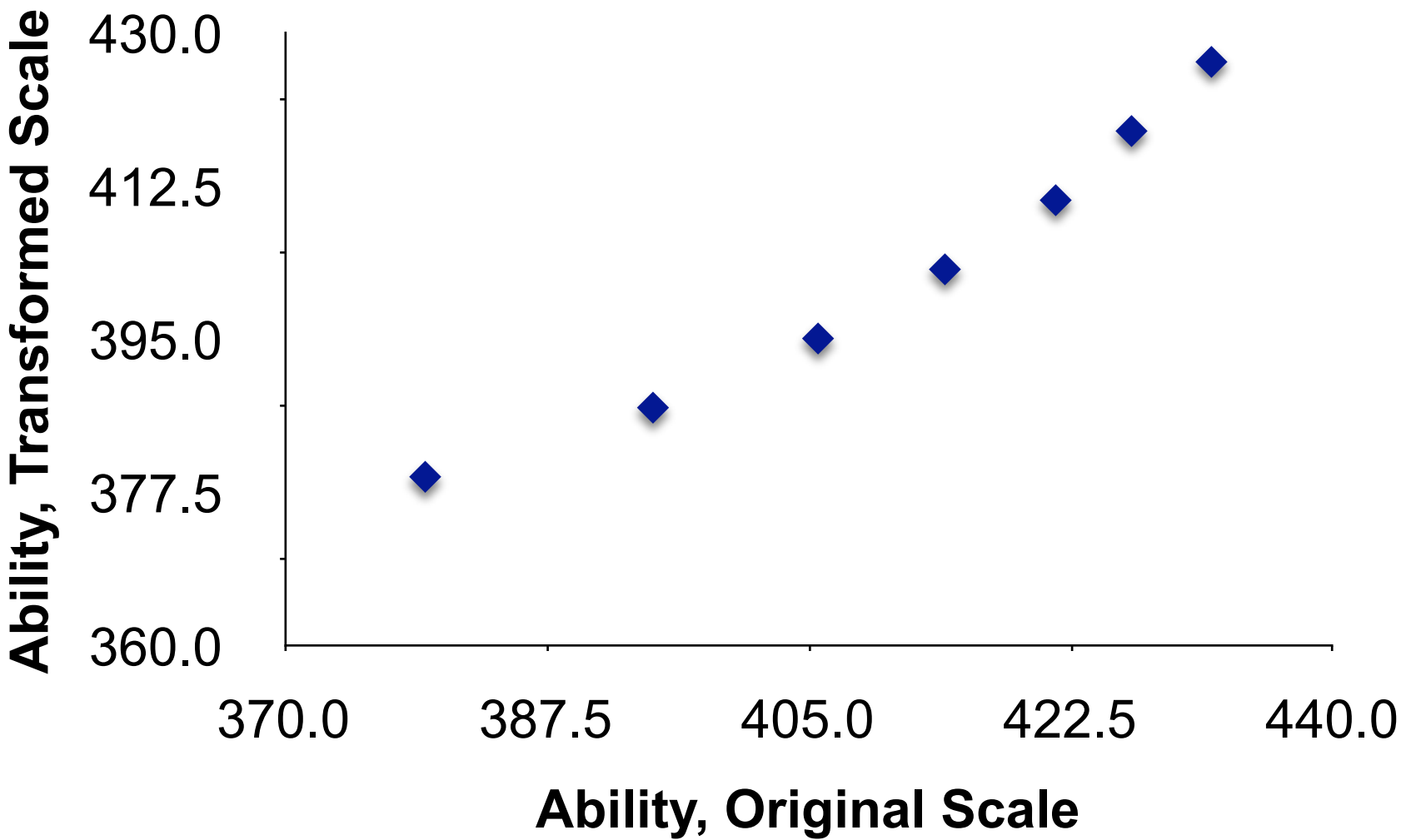
- Bad idea to construct single developmental scale spanning multiple grades
- Even within a single grade, broad range of items required to avoid floor and ceiling effects. Scale compression affects gains of high-achievers vis-à-vis low achievers within a grade.

What to do?

- Use the θ scale anyway, on the assumption that value added estimates are robust to all but “grotesque transformations” of θ .

Test of this hypothesis: rescaled math scores to equate between-grade gains (sample of 19 counties, Southern state, 2005-06)

Figure 5: Median Between-Grade Gains, OriginalScale and Transformed S



Original Scale

Relative to students at the 10th percentile, growth by students at the:

Grade	25th percentile	Median	75 th percentile	90th percentile
2 to 3	0.97	1.06	1.03	0.95
3 to 4	1.10	1.03	1.16	1.34
4 to 5	0.96	1.15	1.35	1.23
5 to 6	1.61	2.12	2.17	2.24
6 to 7	1.21	1.39	1.43	1.62
7 to 8	1.16	1.17	1.16	1.13

Transformed Scale

Relative to students at the 10th percentile, growth by students at the:

Grade	25th percentile	Median	75 th percentile	90th percentile
2 to 3	2.63	5.06	6.83	7.90
3 to 4	1.69	2.10	2.93	3.95
4 to 5	1.30	1.94	2.74	2.87
5 to 6	2.12	3.48	4.25	4.91
6 to 7	1.60	2.29	2.79	3.50
7 to 8	1.51	1.89	2.17	2.35

What to do? (cont.)

- Transform θ to a more acceptable scale $\psi=g(\theta)$ and treat ψ as an interval scale.

Example: normalizing $\Delta\theta$ by mean gain among examinees with same initial score.

Problem: this doesn't produce an interval scale.

What to do? (cont.)

- Map θ to something we can measure on an interval (or even ratio) scale

Examples: inputs, future earnings

What to do? (cont.)

- Ordinal analysis

How it works: Teacher A has n students. Other teachers in a comparison group have m students. There are nm pairwise comparisons. Each comparison that favors Teacher A counts +1 for A. Each comparison that favors the comparison group counts -1 for A. Sum and divide by number of pairwise comparisons.

- Yields an estimate of the probability that a randomly selected student of A outperforms a randomly selected student in the comparison group, minus the probability of the reverse.
- Example of a statistic of concordance/discordance. Somer's d statistic.

- Can control for covariates by conducting pairwise comparisons within groups defined on the basis of a confounding factor (e.g., prior achievement).

Illustration

- Sample of fifth grade mathematics teachers in large Southern city.
- Two measures of value-added
 - Regression model with 5th grade scores regressed on 4th grade scores, with dummy variable for teacher (fixed effect)
 - Somer's d, with students grouped by deciles of prior achievement

Results

- Hypothesis that teachers are ranked the same by both methods rejected ($p=.008$)
- Maximum discrepancy in ranks = 229 (of 237 teachers in all)
- In 10% of cases, discrepancy in ranks is 45 positions or more.
- If teachers in the top quartile are rewarded, more than 1/3 of awardees change depending on which VA measure is used.
- Similar sensitivity in make-up of the bottom quartile.

Conclusions

- It is difficult to substantiate claims that achievement (ability) as measured by IRT is interval-scaled. Strong grounds for skepticism.
- IRT scales appear to be compressed at the high end, affecting within-grade distribution of measured gains.
- Switching to other metrics generally fails to yield an interval- or ratio-scaled measure.
- Ordinal analysis is a feasible alternative.