





Considerations when building and maintaining score scales for reporting test results

Per-Erik Lyrén, Umeå University

Presentation at the 13th SweSAT Conference,
15–17 June, Umeå, Sweden



Scaling Issues

Per-Erik Lyrén, Umeå University

Presentation at the 13th SweSAT Conference,
15–17 June, Umeå, Sweden



Outline

- Why bother?
- Building score scales
 - Recommendations and best practices
- Issues to consider
- Example: The SweSAT Score Scale



Why bother?

- The foundational infrastructure of a testing program
- "The score scale provides the framework for the interpretation of scores. The choice of score scale has implications for test specifications, equating, and test reliability and validity, as well as for test interpretation." (Dorans, 2002)
- Important topic that deserves more attention



Definitions

- *Scaling*:
the process of associating numbers or other ordered indicators with the performance of examinees
- Scaling → a score scale
- Thus, a *score scale* consists of a set of ordered indicators
- How hard can it be to create a score scale??



Examples of score scales

- Admissions tests
 - SAT: 200–800 (400–1600, 600–2400)
 - ACT: 1–36
 - PET: 200–800 (sections 50–150)
 - LSAT: 120–180
 - GMAT: 200–800 (sections 0–60)
 - MCAT: 3–45 (sections 1–15)
- Other examples
 - CAHSEE: 275–450
 - PRAXIS-II: 100–200



Building Score Scales

- Create reported scales that facilitate score meaning and minimize likely misinterpretations and unwarranted inferences (Petersen, Kolen & Hoover, 1989)
 - Score meaning: Normative or content information
 - Misinterpretations: What score scales are there that are used in the same population?



Building Score Scales

- Incorporating normative information
 - Crucial step: Define a norm group
 - Make the raw-to-scale score transformation



Building Score Scales

- Incorporating content information
 - Scale anchoring
 - General statements about what test-takers at different score levels can do
 - Item mapping
 - Representative items for various score points are reported to the test-takers
 - Standard setting



Building Score Scales

- Incorporate score precision information
 - Too few scale score points → loss of precision
 - Too many scale score points → test users might attach significance to small score differences
 - Rule of thumb for the number of score points (see Kolen & Brennan, 2004)
 - Based on test reliability and the choice of confidence interval (CI)



Building Score Scales

- Incorporate score precision information
 - Rule of thumb for the number of score points
 - Example: You want a score scale where a scale score ± 3 score points creates a 68% CI. The reliability of the test is estimated to .91. Then the number of scale score points should be appx. 60. If you want it to create a 95% CI, the number of score points should be appx. 30.



Building Score Scales

- Technical documentation should be made available for users to judge the quality and precision of the scale scores
(from *SEPT*)



Building Score Scales

- Issues to consider:
 - Should all tests within a test battery be scaled the same way?
 - Should estimated true scores be used?
 - Should scores below the chance level be truncated? (For MC tests)
 - If normative information is used in creating the scale, on what reference group should the norms be based?
 - Harris (2007)



Building Score Scales

- Issues to consider:
 - Nonnormative scales (e.g., Angoff, 1962, 1971)
 - "... the meaning that is invested in a scale at the time of its definition is not lasting" and
"... a scale has a reasonable chance of being meaningful to a user if it does not change" (Angoff, 1962, p.32)
 - What makes a scale truly meaningful is *familiarity* and *constancy*.
 - No rescaling required when norms change



Building Score Scales

- Issues to consider:
 - Psychometric models for scaling
 - Thurstone, Guttman, Rasch
 - Etc., etc., etc....



Building Score Scales

- The Well-Aligned Score Scale (Dorans, 2002)
- Seven properties
 - The reference group should be in the middle of the scale
 - Unimodal score distribution for the reference group
 - Nearly symmetric score distribution (about the average score)
 - Shape of distribution should follow a commonly recognized form



Building Score Scales

- Seven properties (cont'd)
 - Working range of scores should extend enough beyond the reported range of scores to permit population shifts
 - The number of scale units should not exceed the number of raw score points
 - A score scale should be viewed as infrastructure that is likely to require repair (e.g. due to shifts in score distribution or when reference groups lose their relevance) → ...



Maintaining score scales

- "Testing programs that attempt to maintain a common scale over time should conduct periodic checks of the stability of the scale on which scores are reported." (Standard 4.17 in *SEPT*)
- Rescale if necessary
- Example: The recentering of the SAT scores in the 1990's



Maintaining score scales

- Equating (multiple test forms)
- Scale drift
 - "... a change in the interpretation that can be validly attached to scores on the score scale" (Haberman & Dorans, 2009)
 - Various reasons: population shifts, content shifts, inconsistent test-construction practices, inadequate anchors for equating.



Maintaining score scales

- Wendler & Walker, 2006 (p.465):
 - How will scale drift be monitored?
 - How frequently is rescaling acceptable?
 - What rescaling method will be used?



Building and Maintaining Score Scales

- Short checklist:
 1. Avoid unwarranted inferences and confusion with other scales
 2. Incorporate score precision information
 3. Incorporate normative and/or content meaning (or not)
 4. Document the scaling process
 5. Maintain the score scale



Example: The SweSAT Score Scale

- Score scale: 0.0–2.0
- Short checklist:
 1. Avoid unwarranted inferences and confusion with other scales
 - Misinterpretations in relation to the upper-secondary school GPA scale (0–20)
 - The 0.0 score: Absence of "ability"?
 2. Incorporate score precision information
 - Using the rule of thumb inversely, 21 score points makes a scale where ± 1 score point corresponds to a 68% CI



Example: The SweSAT Score Scale

- Short checklist:
 3. Incorporate normative and/or content meaning
 - There is normative information, but what is the norm? The reference group has changed
 4. Document the scaling process
 - Poor documentation from when the scale was created



Example: The SweSAT Score Scale

- Short checklist:
 5. Maintain the scale
 - Familiar to most people
 - Has many of the properties of a well-aligned score scale
 - Scale drift?
 - Content and population shifts
 - Shifts in score distributions
 - Original mean: 1.0
 - Current mean: 0.9 (about $\frac{1}{4}$ SD lower)
 - Equating issues
- Conclusion: Change the scale



References and recommended reading

- AERA, APA, & NCME (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- Angoff, W. H. (1962). Scales with nonmeaningful origins and units of measurement. *Educational and Psychological Measurement*, 22(1), 27–34.
- Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., 508–600). Washington, DC: American Council on Education. (Reprinted as W. A. Angoff, *Scales, norms, and equivalent scores*. Princeton, NJ: Educational Testing Service, 1984.)
- Dorans, N. J. (2002). *The recentering of SAT scales and its effects on score distributions and score interpretations* (College Board Research Report No. 2002-11). New York: College Board.
- Dorans, N. J., Pommerich, M., & Holland, P. W. (2007). *Linking and aligning scores and scales*. New York: Springer.



References and recommended reading

- Haberman, S., & Dorans, N. J. (2009). *Scale consistency, drift, stability: Definitions, distinctions and principles*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.
- Harris, D. J. (2007). Practical issues in vertical scaling. In N. J. Dorans et al. (Eds.), *Linking and aligning scores and scales*. New York: Springer.
- Kolen, M. J. (2006). Scaling and norming. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 155–186). Westport, CT: Praeger.
- Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling and linking*. New York: Springer.
- Petersen, N. S., Kolen, M. J., & Hoover, H. D. (1989). Scaling, norming, and equating. In R.L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 221–262). New York: Macmillan.
- Wendler, C. L. W., & Walker, M. E. (2006). Practical issues in designing and maintaining multiple test forms for large-scale programs. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 445–468). Mahwah, NJ: 27 Lawrence Erlbaum Associates.



Thank you for listening!