Variations on a Theme by Thurstone

James Lumsden
The University of Western Australia

A test model based on the Thurstone judgmental model is described. By restricting various parameters of the model, 3 Rasch models, 2 pseudo-Rasch models, 3 two-parameter ICC models, and a Weber's Law model are derived.

The thematic model for latent trait approaches to test scaling was developed by Thurstone (1927) from his law of comparative judgment. The model was variously termed the method of successive intervals (Saffir, 1937), the method of graded dichotomies (Attneave, 1949), and the law of categorical judgment (Torgerson, 1958). In the model, stimuli were conceived as having a discriminal dispersion around a central location on an attribute continuum and judgmental categories (or category boundaries) as having a similar distribution on the same continuum. A variant of this model, which will be called Thurstone Model A, is produced by substituting items for stimuli and persons for judgmental categories, as illustrated in Figure 1.

Thurstone Model A locates items and persons on the same attribute continuum. Each item and each person has a normal distribution of values on the attribute resulting from moment to moment fluctuation. The standard deviations of the distributions may differ from item to item or from person to person. Item and person parameters are assumed to be independent of each other ($r_{ip} = 0$). The correlation between location and dispersion parameters is unspecified for either items or persons. A person answers an item correctly if, at the moment of attempting it, his/her momentary attribute value is higher than the momentary attribute value of the item; otherwise he/she answers it incorrectly.

Torgerson (1958) suggested certain simplifications of the thematic model in order to overcome the formidable estimation problems. He set the standard deviations for categories (subjects) equal and called the result Condition B; he set standard deviations for stimuli (items) equal and called the result Condition C; finally, he set both the standard deviations of categories and the standard deviations of stimuli equal and called the result Condition D. It is the purpose of this paper to examine more systematically a wider range of simplifying possibilities and to relate these to ancient and modern approaches to test scaling.

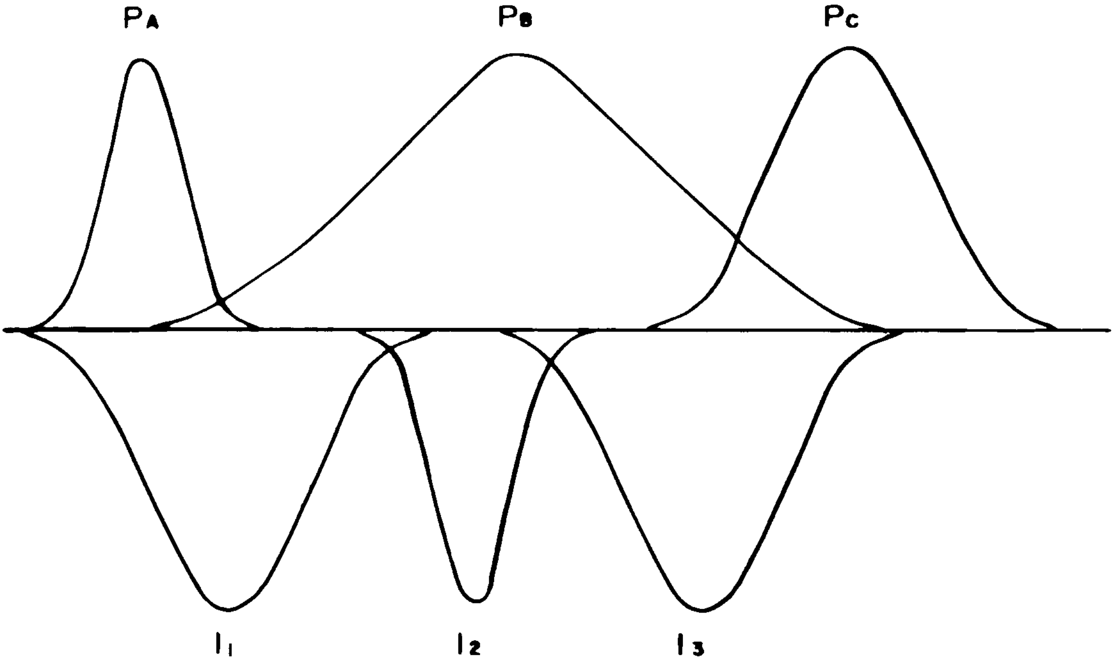
Restrictions on Both Item and Person Dispersions

Torgerson did not consider the cases where item and person dispersions are set to zero. It

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 4, No. 1 Winter 1980 pp. 1–7

Figure 1

Thurstone Model A. Attribute Continuum Showing Locations of Items (I) and Persons (P)



makes no detectable experimental difference whether dispersions for items or persons are set to zero or equal (not zero) except in the degenerate case where both are set to zero, i.e., where both items and persons are perfectly reliable. The latter case is the perfect Guttman scale, which is unlikely to be realized. To preserve heuristic possibilities both zero and equal dispersions will, however, continue to be considered.

Excluding the Guttman scale case, restrictions on both dispersions lead to three Rasch-type models. Rasch 1 has item dispersions set to zero and person dispersions set equal; Rasch 2 has item dispersions set equal and person dispersions set to zero; and Rasch 3 has item dispersions set equal and person dispersions set equal.

From the general logic of the law of comparative judgment, by means set out in Lumsden (1978) or, for a closely analogous case, in Lord and Novick (1968, p. 360), it can be shown that each of these models will generate item characteristic curves (ICCs) which are normal ogives or, what amounts to the same thing, logistics. (The term "logistic" will be used exclusively henceforth but always with the sense "normal ogive or logistic"). All ICCs will have the same slope. Each model will also generate person characteristic curves (PCCs) which are logistics, again with equal slopes. PCCs are plots for each person of the probability of passing against item attribute location (Lumsden, 1977, 1978; Mosier, 1941). Equal slopes for the PCC are required by the Rasch (1960) formulation in order to preserve the specific objectivity of the estimates of person location. No experimental procedure can differentiate the three Rasch-type models; Rasch devotees may prefer the symmetry of Rasch 3. It should be noted that Rasch 3 is formally identical to Torgerson's (1958) Condition D, which in turn is formally identical to Thurstone's (1925) absolute scaling model.

The Rasch model has probably never been realized. Items whose ICCs differ in slope are rejected, and it has been advocated (Andrich, 1973) that persons whose PCCs do not conform to the average slope should be put aside as unmeasurable. One is reminded of Wolfle's (1940, p. 9) famous jibe at the Brown and Stephenson (1933) test of the Spearman two-factor theory: "... if one removes all tetrad differences which do not satisfy the criterion, the remaining ones do satisfy it." However, it needs to be pointed out that estimates made under the Rasch model are quite robust under violations of the assumptions. It is probable that careful attention to test construction will produce items with approximately equal ICC slopes and that variations in PCC slopes will often be so slight as to be negligible in practice.

Restrictions only on the Item Dispersions

If the item dispersions are restricted and person dispersions are permitted to vary, the results are one-parameter ICC models, often incorrectly referred to as Rasch models. Pseudo-Rasch 1 has item dispersions set to zero, and Pseudo-Rasch 2 has item dispersions set equal. For both models there is the additional restriction that person dispersions, while different, are independent of the person locations ($r_{u\sigma} = 0$).

Both the Pseudo-Rasch models will yield logistic ICCs with equal slopes and logistic PCCs with different slopes. Pseudo-Rasch 1 (Lumsden, 1978) has intuitive appeal, since the notion of intrinsic item fluctuation is not plausible for most situations. Pseudo-Rasch 2 is formally identical to Torgerson's (1958) Condition C. The pseudo-Rasch models can never give a poorer, and will generally give a better, fit to data than the limiting case Rasch model. They do this, however, at the cost of a complication that will probably not be acceptable to those who use the Rasch model. Under the pseudo-Rasch models, number correct is no longer a sufficient statistic for estimating the person attribute location (Lumsden, 1977, 1978).

The restriction that location and dispersion for persons be independent, at least for moderate ranges, seems mild (but see below). It would seem that if a unidimensional test can be constructed to strict specifications, then a pseudo-Rasch model will be frequently realized.

Restrictions only on the Person Dispersions

Two-parameter ICC models may be generated by permitting item dispersions to differ subject to the restriction that location and dispersion parameters are independent. Two-Parameter 1 has person dispersions set to zero. Two-Parameter 2 has person dispersions set equal.

These models will yield logistic ICCs with different slopes and logistic PCCs with equal slopes. Two-Parameter 2 is formally identical with Torgerson's Condition B. The models are implausible.

Lumsden (1978) demonstrated that all ICCs must have the same slope if items are strictly unidimensional, defined as systematically measuring the same thing. Lord and Novick (1968, chap. 16) avoided this argument by postulating a weakly unidimensional test in which the items share a single common factor but with each item also having a specific factor orthogonal to the common factor and to all other specific factors. Variation in the relative effect of the specific factor variance will produce differences in ICC slope. Such items will meet the unit rank criterion, but the "specifics" are only specific within the test space. Selection of further items from the general item domain will show that the large specifics are really group factors.

It should be noted that if there are large specific effects, then the ICCs will not be sample free. It will be possible, in principle, to find groups who differ in their distributions on the specific factor and their ICCs will, in consequence, differ in slope (a_g) or location (b_g) or both.

The worst case for weak unidimensionality is a collection of diverse items sharing no clearly definable common factor. Thus, a test composed of one item from each of the Primary Mental Abilities tests may yield unit rank; selection of another item will lead to a breakdown of local independence.

The best case is given by a test where the items have a definable common factor, but some have large specifics. A sensibly constructed vocabulary test will sample items from special subvocabularies so that only one item from each is included in the final test. Tests for unit rank or local independence will ensure orthogonality of the specifics, but these are rarely applied. But suppose now that a test is made for children aged 10 requiring the understanding of two words: *bassoon* and *bunt. Bassoon* is easier for children who have studied music and probably for girls; *bunt* is easier for baseball fans and probably for boys. The test is ridiculously biased as a measure of general verbal comprehension. If the test is long, the effect of the mutually orthogonal specifics becomes less, and some justification is given for the two-parameter model. But the large specific variances are noise in this context. Is it not a sensible plan to eliminate items with large specifics? This returns us effectively to the one-parameter case: negligible specifics and items that measure the same thing in an unbiased way.

It is ironic that adaptive testing procedures developed for the two- and three-parameter models unwittingly approximate the one-parameter case. Selection of items by maximum information methods implies that the low slope items are rarely, in some cases never, used. Effectively, the item bank consists only of the high slope items.

The two-parameter models have been extensively used (see Bock & Wood, 1971; Lord & Novick, 1968; Lumsden, 1976). Attention has been focused exclusively on scaling from the ICCs, and the PCCs have not been considered. It is significant that no study has yet demonstrated any decisive advantage for the weighted procedures based on these models over simpler unweighted procedures (see, for example, Dinero & Haertel, 1977).

Restrictions only on Dependency of Location and Dispersion

Another two-parameter model is generated if both item dispersions and person dispersions are permitted to differ, subject only to the restriction that location and dispersion be independent. For this model, Two-Parameter 3, all ICCs will be logistics with different slopes and all PCCs will be logistics with different slopes. This model does not appear to have been studied; but it seems that if a two-parameter model can be realized at all, it will be this one rather than the limiting cases, Two-Parameter 1 and 2.

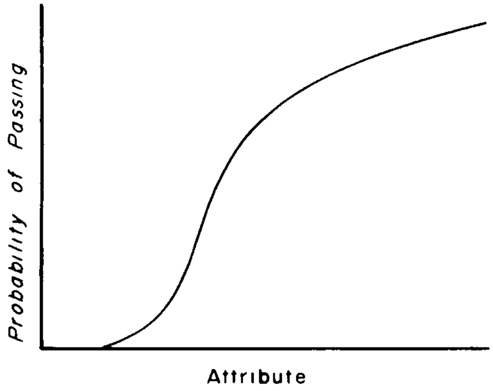
Relaxing the Independence Restriction—Weber's Law

Relaxation of the independence restriction produces many variations of Model A, which have been little studied. Consideration will be restricted to one simple example. Suppose that person dispersions are permitted to differ so that the average dispersion at any attribute level is a linear function of the person location ($r_{u\sigma} > o$). Further simplification is obtained by setting item dispersions to zero. This will produce a model for which Weber's Law will hold. Persons with higher attribute locations will have proportionally greater dispersions than persons with lower attribute locations, and it will therefore be more difficult for the test to discriminate.

For the Weber's Law model, ICCs will not be logistic but eccentric S-shaped curves in which the positive acceleration of the slope below the item location value is greater than the absolute value of the negative acceleration of the slope above the item location value (Figure 2). It is probable, however, that in practice, ICCs de-

Figure 2

Item Characteristic Curve for the Weber's Law Model (Eccentricity Probably Exaggerated)



rived from the Weber's Law model will differ only slightly, and usually undetectably, from logistic ICCs.

A more promising test of the model may be derived from the PCCs which will be logistics differing in slope. The average slopes of the PCCs will be an inverse function of person attribute location. Average standard deviations can be estimated from the slopes; and if Weber's Law holds, the plot of average standard deviation against attribute location will be linear (Figure 3).

Zero Point of Ability

Thurstone (1928) pointed out that as the location parameter approaches the absolute zero, the dispersion parameter approaches zero, since negative values for the attribute are impossible. It follows that if attribute location can be measured on an interval scale and if Weber's Law holds, then a ratio scale is achievable. Thurstone noted that the standard deviation of mental ages was a linear function of average mental age, i.e., of age. (Thurstone did not appear to have realized that he was asserting Weber's Law). He extrapolated the plot to zero standard deviation and found, happily, that the zero point for intelligence came some months before birth. Unfortunately, the Thurstone result is an artifact of intelligence test construction. The standard deviation of mental ages is deliberately increased with age in order to keep IQs approximately constant. Experienced testers say that they feel more confident of their discriminations with dull children than with bright children, but this may be an aspect of the same artifact.

An implication of the Weber's Law model is that the slope of the ICC will be negatively correlated with location. Lord (1975) found, however, that for items from a multiple-choice mathematical aptitude test, there was a significant positive correlation between slope and location. It was shown that this was unlikely to be an artifact of the estimation procedure. Urry (1974, Table 3) provided a bivariate distribution of slope and location for 200 items of a verbal ability item bank. There was no evidence for either a positive or negative relationship. Similar results were obtained for data supplied by Weiss (1973) for an adaptive vocabulary test. These results cannot be regarded as completely cogent, since the unidimensionality requirement was not met; but they are not at all encouraging.

Figure 3

Plot of Average Standard Deviation Against Location for Weber's Law Model

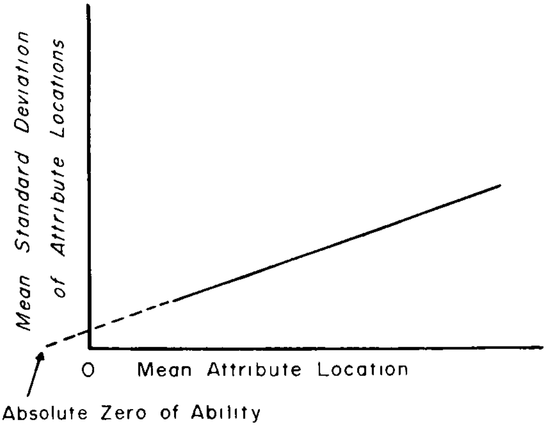


Table 1

Variants of Thurstone Model A under Various Restrictions

Restrictions				Model	ICC		PCC		Comments on applicability
Item		Person							
$r_{u\sigma}$	σ	$r_{u\sigma}$	σ		Form	Slopes	Form	Slopes	
-	Zero	- -	Zero	Guttman Scale	Quantum	Infinite	Quantum	Infinite	Not realizable
-	Zero	-	Equal	Rasch 1	Logistic	Equal	Logistic	Equal	Realizable
-	Equal	-	Zero	Rasch 2	Logistic	Equal	Logistic	Equal	(approximately)
-	Equal	-	Equal	Rasch 3	Logistic	Equal	Logistic	Equal	
-	Zero	Zero	Different	Pseudo-Rasch 1	Logistic	Equal	Logistic	Different	Realizable
-	Equal	Zero	Different	Pseudo-Rasch 2	Logistic	Equal	Logistic	Different	
Zero	Different	-	Zero	Two-parameter 1	Logistic	Different	Logistic	Equal	Not realizable
Zero	Different	-	Equal	Two-parameter 2	Logistic	Different	Logistic	Equal	
Zero	Different	Zero	Different	Two-parameter 3	Logistic	Different	Logistic	Different	Not realizable
-	Zero	Greater than Zero	Different	Weber's Law	Eccentric S	Different	Logistic	Different	? ?

Conclusion

The principal results of the discussion of scaling models are given in Table 1. It is difficult not to be impressed with the power and versatility of Model A. During the 1920s Thurstone stole fire from the gods. (As a punishment they chained him to factor analysis.) It is true that many modern theorists prefer to work with less primitive models and can argue cogently that it is not necessary to consider concepts underlying ICCs, PCCs, and the like. The argument is, however, shortsighted. It can do no harm and may do much good to take advantage of the heuristic power of the models to provide "freeing moves" in an impasse.

References

Andrich, D. *Latent trait psychometric theory in the measurement and evaluatation of essay-writing ability.* Unpublished doctoral dissertation, University of Chicago, 1973.

Attneave, F. A method of graded dichotomies for the scaling of judgments. *Psychological Review,* 1949, *56,* 334-340.

Bock, R. D., & Wood R. Test theory. *Annual Review of Psychology,* 1971, *22,* 193-224.

Brown, W., & Stephenson, W. A test of the theory of two factors. *British Journal of Psychology,* 1933, *23,* 352-370.

Dinero, T. E., & Haertel, E. Applicability of the Rasch model with varying item discriminations. *Applied Psychological Measurement,* 1977, *1,* 581-592.

Lord, F. M. The 'ability' scale in item characteristic curve theory. *Psychometrika,* 1975, *40,* 205-217.

Lord, F. M., & Novick, M. R. *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley, 1968.

Lumsden, J. Test theory. *Annual Review of Psychology,* 1976, *27,* 251-280.

Lumsden, J. Person reliability. *Applied Psychological Measurement,* 1977, *1,* 477-482.

Lumsden, J. Tests are perfectly reliable. *British Journal of Mathematical and Statistical Psychology,* 1978, *31,* 19-26.

Mosier, C. I. Psychophysics and mental test theory II. The constant process. *Psychological Review,* 1941, *48,* 235-249.

Rasch, G. *Probabalistic models for some intelligence and attainment tests.* Copenhagen, Denmark: Danmarks Paedogogiske Institut, 1960.

Saffir, M. A comparative study of scales constructed by three psychophysical models. *Psychometrika,* 1937, *2,* 179-198.

Thurstone, L. L. A method of scaling psychological and educational tests. *Journal of Educational Psychology,* 1925, *16,* 433-451.

Thurstone, L. L. A law of comparative judgment. *Psychological Review,* 1927, *34,* 273-286.

Thurstone, L. L. The absolute zero in intelligence measurement. *Psychological Review,* 1928, *35,* 175-197.

Torgerson, W. S. *Theory and methods of scaling.* New York: John Wiley, 1958.

Urry, V. W. *Computer-assisted testing: Calibration and evaluation of the verbal ability bank* (TS-74-3). Washington, DC: U.S. Civil Service Commission, Personnel Research and Development Center, 1974.

Weiss, D. J. *The stratified adaptive computerized test* (Research Report 73-3). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, September 1973. (NTIS No. AD 768376)

Wolfle, D. *Factor analysis to 1940.* Psychometric Monograph, 1940, (No. 3).