Measurement and Personality Assessment
Edw. E. Roskam (ed.)

ADDITIVITY IN PSYCHOLOGICAL MEASUREMENT

Benjamin D. Wright
MESA Psychometric Laboratory
The Department of Education
The University of Chicago

Shows how the Rasch model is a unit-maintaining process (Thurstone, 1931) which enables the construction of additivity (Campbell, 1920) and hence fundamental measurement (Luce and Tukey, 1964). Provides the basic statistics for determining the extent to which additivity has been approximated with particular data. (A note reviews the obstacles to maintaining units or constructing additivity encountered by binomial response models with more than one item parameter.)

Key Words: additivity, conjoint measurement, fit, fundamental measurement, interval scales, item response theory, latent trait theory, one parameter models, Rasch models.

INTRODUCTION

The realization that additivity can be constructed for psychological research is often traced to Luce and Tukey (1964). They show that a conjoint additivity as good for measuring as that produced by physical concatenation can be obtained from responses produced by the interaction of two kinds of objects (e.g., persons and test items). All that is necessary is that the interaction be conducted so that its outcomes (e.g., the persons' responses to the items) are dominated by a linear combination of two kinds of quantities (e.g., person measures and item calibrations).

Thurstone's 1927 Law of Comparative Judgement contains the same idea (Andrich, 1978) and his empirical work of 1928, 1929 and 1931 provides rough examples of additivity. The construction of additivity also occurs in Bradley and Terry (1952) and Rasch (1958, 1960, 1966).

The additivity which follows from Rasch's "specific objectivity" is developed in Rasch 1960, 1961, 1967 and 1977. Specific objectivity and estimation sufficiency are two sides of the same approach to inference, i.e., that the statistical model on which inference is to be based be factorable in its parameters. Andersen (1977) shows that the only response processes which support specific objectivity and hence additivity are those which have sufficient statistics for their parameters.

Several authors find additive conjoint measurement in Rasch's work (Keats, 1967; Fischer, 1968; Brogden, 1977). Perline, Wright and Wainer (1977) provide two examples of the extent to which the Rasch process can organize data so that they satisfy the monotonicity and double cancellation requirements of conjoint measurement. Wright and Stone (1979) show how to obtain additivity from mental tests. Wright and Masters (1982) give examples of the construction of additivity from rating scale and partial credit data.

MAINTAINING A UNIT

"All measurement implies the recreation or restatement of the attribute measured to an abstract linear form. . . . A unit of measurement is always a process of some kind which can be repeated without modification in the different parts of the measurement continuum" (Thurstone, 1931, 257).

Rasch (1960, 171-172) shows that, if

$$P = \exp(b - d)/G$$
$$G = 1 + \exp(b - d)$$

is the way person ability b and item difficulty d combine to govern the probability of a successful outcome and, if Event AB is person A succeeding but person B failing on some item, while Event BA is person B succeeding but person A failing on the same item, a distance between persons A and B on a scale defined by a set of items of a single kind can be estimated by

$$b_A - b_B = \log N_{AB} - \log N_{BA}$$

where N_{AB} is the number of times A succeeds but B fails and N_{BA} is the number of times B succeeds but A fails on any set of these items.

This happens because,

$$P_{AB} = P_A(1 - P_B) = \exp(b_A - d)/G_A G_B$$

and

$$P_{BA} = P_B(1 - P_A) = \exp(b_B - d)/G_A G_B$$

so that d cancels out of the odds for Event AB over Event BA

$$P_{AB}/P_{BA} = \exp(b_A - b_B)$$

causing the log odds (logit)

$$\log(P_{AB}/P_{BA}) = b_A - b_B$$

to be a distance which holds regardless of the value of d. This makes Rasch's model for specifying measures a unit-maintaining process of the kind Thurstone requires.

CONSTRUCTING ADDITIVITY

Campbell (1920) identifies additivity as the hallmark of measurement. The way to construct additivity for psychologocal measurement is to devise an operation which answers the question: "If person A has more ability than person B, then how much 'ability' must be <u>added</u> to B to make the <u>performance</u> of B <u>appear the same</u> as the performance of A?"

To answer this we review how ability becomes known. In order to observe the abilities of persons A and B we must expose them to situations which provoke manifestations of their ability. This narrows the question to: "What change in the situation through which we find out about person ability, say by testing persons with items, will give B the same probability of success as A?" To be specific, "What item j will make the performance of person B appear the same as the performance of person A on item i?"

According to the Rasch process, the way to get $P_{Bj} = P_{Ai}$ is to make $b_B - d_j = b_A - d_i$.

The 'addition' required to cause B to perform like A is

$$b_B + (b_A - b_B) = b_A .$$

The way to perform this 'addition' is to test person B with an item j

of difficulty

$$d_j = d_i + (b_B - b_A) \quad .$$

The way to evaluate the quality of this 'addition' is to observe the extent to which the performance of person B on items like j is statistically equivalent to the performance of person A on items like i . This is the kind of equivalence which is checked when response residuals are analyzed for their fit to the Rasch process.

GUIDING THE CONSTRUCTION OF ADDITIVITY BY ANALYZING FIT

In order to go forward with the construction of additivity, we need a way to evaluate how well we are doing at each step. We need to know the extent to which the arithmetic we plan to do with our measures will hold up.

The best way to evaluate the extent of additivity (i.e., scale invariance) obtained by the Rasch process from a particular set of data is to compose a score residual $y = x - Ex$ for each response x and then to accumulate these score residuals and their squares over the item-person response subsets for which scale invariance is suspect. Response subsets can be defined by any combination of items and persons which might interact in a way that disturbs additivity.

The expected response Ex is estimated from the current Rasch estimates of person ability b and item difficulty d . (For binomial data $x = 0$ or 1 , Ex is the probability, $P = \exp(b - d)/(1 + \exp(b - d))$. For comparable statistics for rating scale, partial credit, repeated trial and Poisson data see Wright and Masters, 1982, 100).

If we let $(f_1 - f_0)$ represent the extent to which a particular subset of responses fails to maintain the additivity implied by the majority of items and persons, then the sum of score residuals for that subset, Σy , estimates

$$(f_1 - f_0)\,\Sigma(dy/df) \quad .$$

The differential of y with respect to f

$$dy/df = dE/df = Vx = w$$

is the parameter information in the observed response and also the score variance and the inverse of the logit variance. (For the binomial case $dy/df = dP/df = P(1 - P) = w$.)

We can use $\Sigma y \approx (f_1 - f_0)\,\Sigma w$

to form $(f_1 - f_0) \approx \Sigma y/\Sigma w = g$ so that

the BIAS $g = \Sigma y/\Sigma w$

estimates the logit discrepancy in scale invariance $(f_1 - f_0)$ associated with the response subset specified.

The noise within a response subset can be evaluated by comparing the observed squared residual y^2 with its expectation w .

The mean square <u>standardized</u> residual,

$$u = \Sigma(y^2/w)/\Sigma 1 = \Sigma z^2/\Sigma 1$$

is sensitive to unexpected responses when $(b - d)$ is absolutely large because w diminishes exponentially as the distance between b and d increases. This makes u useful for detecting episodic outliers like lucky guesses and careless mistakes.

The mean square <u>information</u> weighted residual,

$$v = \Sigma w z^2/\Sigma w = \Sigma(w y^2/w)/\Sigma w = \Sigma y^2/\Sigma w$$

focuses on responses from proximate b and d which contribute most to their estimation. This makes v useful for detecting systematic disturbances like loss of local independence and loss of unidimensionality.

Values of u and v substantially greater than one signal disruptions in additivity of the kind caused by ambiguities and errors in task presentation, response representation, recording and scoring. Values substantially less than one signal loss of independence of the kind caused by systematic omissions, item confounding, person collusion, prior exposure and curriculum interaction.

When data approximate the Rasch process, the expectations and variances of these fit statistics can be represented closely enough by $Eg = 0$, $Vg = 1/\Sigma w$, $Eu = Ev = 1$ and $Vu = Vv = 2/\Sigma 1$ to provide a frame of reference for supervising the construction of additivity.

This representation can be improved by dividing g and by multiplying u and v by a factor which corrects for the use of parameter estimates in the calculation of response expectations Ex. The factor is obtained by dividing the total number of responses x in the subset by the degrees of freedom which remain after the number of parameter estimates needed to calculate the corresponding Ex has been deducted.

It is convenient to work with cube root standardizations of u and v (Wright and Masters, 1982, 100) referred to as:

OUTFIT for $g(u)$, because it detects <u>out</u>liers in the <u>out</u>er regions of person-item interactions where $(b - d)$ is absolutely large, and

INFIT for $g(v)$, because it is weighted by the parameter <u>in</u>formation borne by response x and evaluates the <u>in</u>ner region of person-item interactions where $(b - d)$ is absolutely small.

CONCLUSIONS

It has long been customary in psychological research to construct scores by counting answers (scored by their ordinal position in a sequence of ordered response possibilities) and then to use these scores (and monotonic transformations of them) as measures. When the questions asked have only two answer categories, we count right answers. When the questions offer an ordered series of answer categories, we count how many categories from 'least' to 'most' ('worst' to 'best', 'weakest' to 'strongest') have been surpassed.

If there has been any progress in quantitative psychology, this kind of counting must have been useful. This has implications. Counting this way implies a particular measurement process. Counting implies a process which derives counting as the necessary and sufficient scoring procedure.

Whether particular data can be organized to follow the Rasch process can only be discovered by applying the process and examining the consequences. It is worth noticing, however, that whenever we have deemed it useful to count right answers (as in educational testing) or to add scale ratings (as

in Likert scaling), we have taken it for granted that the data concerned did, in fact, follow a process identical to the Rasch process well enough to suit our purposes. This is because the Rasch process is the only response process for which counts and additions are the sufficient statistics.

Since the Rasch process constructs conjoint additivity whenever data are valid for such a construction, we have, in our counting, been taking the first steps toward additivity all along. All we need do now is to take this implication of our actions seriously and to complete our data analyses by verifying the extent to which our data fit the Rasch process.

If we subscribe to Thurstone's and Campbell's requirements for measurement, then fitting the Rasch process becomes more than a convenience, it becomes the essential criterion for data good enough to support the construction of additivity. When data can be organized to fit well enough to be useful, then we can use the results to define Thurstone linear scales and to make Luce and Tukey fundamental measures on them.

<u>Note concerning the failure of binomial response processes with two and three item parameters to maintain units or enable the construction of additivity</u>.

Consider the three item parameter binomial process

$$Q = c + (1 - c)P \qquad P = \exp(a(b - d))/G$$
$$1 - Q = (1 - c)(1 - P) \qquad G = 1 + \exp(a(b - d))$$

and form the odds for Event AB over Event BA as before,

$$Q_{AB}/Q_{BA} = Q_A(1 - Q_B)/Q_B(1 - Q_A)$$
$$= \frac{c(1 - P_B) + (1 - c)P_A(1 - P_B)}{c(1 - P_A) + (1 - c)P_B(1 - P_A)}$$

If all three item parameters remain variable, there is no way to cancel any of them out of this expression in order to maintain a unit among b's over the ranges of the item parameters. There is also no way to cancel b out of this expression in order to enable a sample-free estimation of any of the item parameters.

If we make c a known constant, always the same for all items and persons no matter how much persons differ in their guessing behavior, we could use

$$\frac{Q - c}{1 - Q} = \frac{P}{1 - P} = \exp(a(b - d))$$

to eliminate the influence of this one common c and concentrate on the problems caused by the interaction of b with a. But when c varies from item to item, then, even when its values are known, the differential consequences of b variation on

$$(c/(1 - c))(1 - P_B) \quad \text{versus} \quad (c/(1 - c)(1 - P_A)$$

prevent the Q process from maintaining a fixed distance between persons A and B over the range of d and c.

Nor can we construct an addition for the Q process. There is no fixed amount which, when 'added' to b_B, will make $Q_{Bj} = Q_{Ai}$ so that the performance of person B can become stochastically equivalent to the performance of person A. The amount to add necessarily varies with the varying values of c and a.

If we abandon c as a variable, and focus on a response model with two item parameters, then

$$P_{AB}/P_{BA} = \exp(a(b_A - d))/\exp(a(b_B - d))$$

and

$$\log(P_{AB}/P_{BA}) = a(b_A - b_B) \quad .$$

The item parameter d is gone, so that $a(b_A - b_B)$ is maintained over the range of d. But what shall we do if parameter a is allowed to vary?

If we advance a as a second item parameter, we have to estimate a different unit for every item. The distance between A and B can only be maintained if every a for every item can be known independently of every b to be compared. That prevents us from using the behavior of persons to estimate the values of a. This happens because when we try to estimate a we find that we cannot separate it from its interactions with the estimation of the b's used for its estimation. When we try to estimate these b's we find that we cannot separate them from their interactions with a. (Advancing a as a second person parameter runs into the same kind of trouble but with d instead of b.)

We can maintain the distance between A and B only when a is a constant over persons and items, that is, when we are back to the Rasch process.

Nor can the process which includes a as a variable support additivity.

When $\quad P = \exp(a(b - d))/(1 + \exp(a(b - d)))$

then $\quad P_{Bj} = P_{Ai}$

implies that $\quad a_j(b_B - d_j) = a_i(b_A - d_i)$

so that $\quad b_A = d_i + (a_j/a_i)(b_B - d_j)$

An 'addition' which will equate the performances of persons A and B is uniquely defined only over persons and items for which a is a constant so

that $\quad (a_j/a_i) = 1$

and $\quad b_A - b_B = (d_i - d_j)$

as in the Rasch process.

If measurement is our aim, nothing can be gained by chasing after extra item (or person) parameters like c and a. We must seek, instead, for items which can be managed by an observation process in which any potentially misleading disturbances are kept slight enough to preserve the necessary scale stability.

REFERENCES

Andersen, E.B. (1975). Sufficient statistics and latent trait models. Psychometrika, 42, 69-81.

Andrich, D. (1978). Relationships between the Thurstone and Rasch approaches to item scaling. Applied Psychological Measurement, 3, 449-460.

Bradley, R.A. and Terry, M.E. (1952). Rank analysis of incomplete block designs I: The method of paired comparisons. Biometrika, 39, 324-345.

Brogden, H.E. (1977). The Rasch model, the law of comparative judgement and additive conjoint measurement. Psychometrika, 42, 631-634.

Campbell, N.R. (1920). *Physics: The elements*. London: Cambridge University Press.

Fischer, G. (1968). *Psychologische testtheorie*. Bern: Huber.

Keats, J.A. (1967). Test theory. *Annual Review of Psychology*, *18*, 217-238.

Luce, R.D. and Tukey, J.W. (1964). Simultaneous conjoint measurement: A new type of fundamental measurement. *Journal of Mathematical Psychology*, *1*, 1-27.

Perline, R., Wright, B.D. and Wainer, H. (1979). The Rasch model as additive conjoint measurement. *Applied Psychological Measurement*, *3*, 237-256.

Rasch, G. (1958). *On applying a general measuring theory of bridgebuilding between similar psychological tests*. Copenhagen: Danmarks Paedagogiske Institut.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danmarks Paedogogiské Institut. (Chicago: University of Chicago Press, 1980).

Rasch, G. (1961). On general laws and the meaning of measurement in psychology. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, *4*, 321-323.

Rasch, G. (1966). An individualistic approach to item analysis. In P.F. Lazersfeld and N.W. Henry (Eds.), *Readings in mathematical social science*. Chicago: Science Research Associates.

Rasch, G. (1967). An informal report on the present state of a theory of objectivity in comparisons. In L.J. van der Kamp and C.A.J. Viek (Eds.), *Proceedings of the NUFFIC International Summer Session in Science at "Het Oude Hof."* Leiden.

Rasch, G. (1977). On specific objectivity: An attempt at formalizing the request for generality and vaildity of scientific statements. *Danish Yearbook of Philosophy*, *14*, 58-94.

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

Thurstone, L.L. (1928). Attitudes can be measured. *American Journal of Sociology*, *33*, 529-554.

Thurstone, L.L. (1929). Theory of attitude measurement. *Psychological Review*, *36*, 222-241.

Thurstone, L.L. (1931). The influence of motion pictures on children's attitudes. *Journal of Social Psychology*, *2*, 291-305.

Wright, B.D. and Masters, G.N. (1982). *Rating scale analysis*. Chicago: MESA Press.

Wright, B.D. and Stone, M.H. (1979). *Best test design*. Chicago: MESA Press.