

Relationships Between the Thurstone and Rasch Approaches to Item Scaling

David Andrich
The University of Western Australia

When the logistic function is substituted for the normal, Thurstone's Case V specialization of the law of comparative judgment for paired comparison responses gives an identical equation for the estimation of item scale values as does the Rasch formulation for direct responses. The law of comparative judgment must be modified to include a subject parameter; but this parameter, which is

eliminated statistically with respect to the direct response design, is eliminated experimentally in the paired comparison design. Some comparisons and contrasts are made between the two approaches to item scaling, and it is shown that greater generalizability for item scaling is possible when the two approaches are juxtaposed appropriately.

The application of Thurstone's law of comparative judgment to items which have no physical properties (e.g., statements), permits identification of a conformable set of items which can then be used to measure subjects on a psychological continuum. More specifically, both the evidence that a set of items does refer to a unidimensional continuum and the relative scale values of the items are provided.

The basic data collection design for scaling items in the Thurstone tradition is that of paired comparisons, while the most common design for measurement requires subjects to respond to items according to some positive-negative dichotomy, such as "agree" or "disagree." To contrast this latter design with that of paired comparisons, it will be called the *direct response* design. Measurements of subjects are then usually the medians of the scale values of the items they endorse positively (Edwards, 1957).

An alternate approach to scaling items and measuring subjects is available from latent trait psychometric theory. Subjects who are to be measured respond according to the direct response design, but these same responses may be used both to scale items and to measure subjects. A number of statistical models have been proposed for transforming the responses to these two ends. One of the simplest and most elegant ones is the Rasch (1960, 1961) simple logistic model (SLM) which has been extensively studied by Wright and others (Wright, 1967; Wright & Panchapakesan, 1969; Wright & Douglas, 1978) and by Andersen (1972a, 1973a, 1973b).

Wainer, Fairbank, and Hough (1978) report a study in which the scale value of items obtained from a paired comparison design and analyzed according to Thurstone's Case V specialization for the

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 2, No. 3 Summer 1978 pp. 449-460
© Copyright 1978 West Publishing Co.

law of comparative judgment were equivalent (after an adjustment for the arbitrary origin and unit) to the scale values of the same items obtained from a direct response design and analyzed according to the SLM. Wainer et al. go on to say, "The measurement model underlying both data collection schemes was very similar in that the Rasch Model posits a logistic response function with equal slopes, as does the Bradley-Terry-Luce model used for the scale values" (pp. 318-319).

The present paper explains in detail the similarities and differences between applications of Thurstone's Case V model, which is numerically equivalent to the Bradley-Terry-Luce model, and the SLM. It is shown that in a well-defined sense, item scale values obtained from the two approaches should be equivalent. The Thurstone model needs to be qualified to include the subject parameter contained in the SLM model. However, a critical point in unifying the approaches is that this parameter, which can be eliminated explicitly by statistical conditioning in Rasch's model, is also eliminated in the paired comparison design.

The Thurstone Formulation

The results derived below are well known and follow Thurstone (1927a) and Bock and Jones (1968). They are presented in order to motivate the required modification, which, interestingly, is consistent with Thurstone's own statements. It is stressed, however, that the concern is not with detailed elaborations of Thurstone's formulations, as found in Torgerson (1958), for example. The prime concern is with basic principles; therefore, only certain limited features relevant to the discussion are highlighted.

The Law of Comparative Judgment

In the paired comparison design, subjects compare items in pairs and report which item they judge to reflect more of the trait in question. This comparison is formalized as follows:

1. When person v encounters item j , he/she perceives it to have a value D_{vj} on the trait. D_{vj} , called the discriminial process, is a continuous random variable defined by

$$D_{vj} = \alpha_j + \epsilon'_{vj}, \quad [1]$$

where α_j is the scale value of item j and is common to it with respect to all subjects and ϵ'_{vj} is the specific error component associated with subject v .

Over a population of subjects, D_{vj} is assumed normally distributed with mean α_j and variance σ_j^2 . Therefore, ϵ'_{vj} has a mean of zero and variance σ_j^2 in the population of subjects.

2. When subject v compares two statements, j and i , he/she reports that item j reflects more of the property of $D_{vj} - D_{vi} > 0$. Since D_{vj} and D_{vi} are assumed normally distributed, the difference process

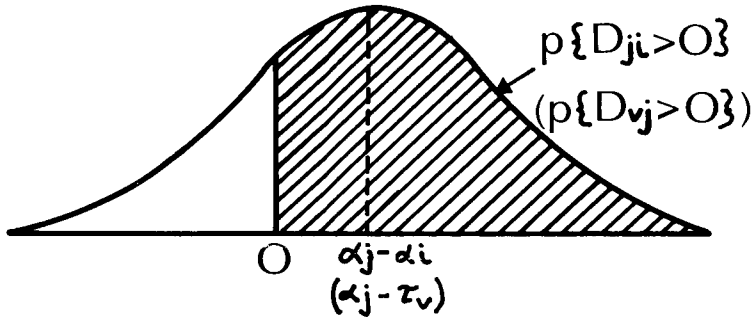
$$D_{ji} = D_{vj} - D_{vi} = (\alpha_j - \alpha_i) + (\epsilon'_{vj} - \epsilon'_{vi}) \quad [2]$$

is normally distributed with mean $\alpha_j - \alpha_i$ and variance σ_{ji}^2 given by

$$\sigma_{ji}^2 = \sigma_j^2 + \sigma_i^2 - 2\rho_{ji}'\sigma_j'\sigma_i' \quad [3]$$

The assumed distribution of D_{ji} is depicted in Figure 1, where the shaded region represents the probability that D_{ji} is positive.

FIGURE 1: Distribution of D_{ji} (D_{vj}) showing the probability that D_{ji} (D_{vj}) is positive.



Let the random variable x_{ji} take the value 1 when $D_{ji} > 0$, and 0 otherwise. Then,

$$p\{x_{ji} = 1 | \alpha_j, \alpha_i, \sigma_{ji}^2\} = \frac{1}{\sigma_{ji} \sqrt{2\pi}} \int_0^{\infty} \exp\left[-\frac{1}{2} \frac{[t - (\alpha_j - \alpha_i)]^2}{\sigma_{ji}^2}\right] dt. \quad [4]$$

Introducing a Subject Parameter

Instead of proceeding directly from the above, some important points Thurstone made with respect to the correlation ρ'_{ji} should first be examined closely.

If you look at two handwriting specimens in a mood slightly more generous and tolerant than ordinarily, you may perceive a degree of excellence in specimen A a little higher than its mean excellence. But at the same moment specimen B is also judged a little higher than its average or mean excellence for the same reason. To the extent that such a factor is at work the discriminational derivations will tend to vary together and the correlation r will be high and positive. (Thurstone, 1927a, p. 279)

The above statements are now formalized by introducing a subject parameter τ_v for the “mood” of subject, or observer, v . It is clear that Thurstone considered this effect to be part of the error. Accordingly, ϵ'_{vj} is qualified by $\epsilon'_{vj} = f(\tau_v, \epsilon_{vj})$, where ϵ_{vj} represents only an error with respect to, or within, a given subject.

Thurstone’s intuitions about the effect of τ_v suggests that it would inflate or deflate D_{vj} systematically across all objects. For convenience in further exposition, but without loss of generality, this effect is expressed in the negative form. Therefore, let $f(\tau_v, \epsilon_{vj}) = -\tau_v + \epsilon_{vj}$, giving

$$D_{vj} = \alpha_j - \tau_v + \epsilon_{vj}. \quad [5]$$

For a given subject, $E[D_{vj} | \tau_v] = \alpha_j - \tau_v$ and $V[D_{vj} | \tau_v] = \sigma_j^2$. With respect to a population of subjects, $E[D_{vj}] = \alpha_j$ and $V[D_{vj}] = \sigma_j^2$ still holds, since $E[\tau_v] = 0$ is assumed. Consequently, if the within-subject-error is independently and homogeneously distributed with respect to subjects and items,

$$v[D_{vj}] = \sigma_j'^2 = \sigma_\tau^2 + \sigma_j^2, \quad [5a]$$

where σ_τ^2 is the variance of τ 's in the population of subjects. Expressing it in words, in Equation 1, in contrast to Equation 5, the variance among subjects, σ_τ^2 , is absorbed with the error variance within individual discriminial processes σ_j^2 to give $\sigma_j'^2$.

Consider now the consequence of τ_v for the difference process between j and i . The discriminial process for subject v and item i is

$$D_{vi} = \alpha_i - \tau_v + \epsilon_{vi}, \quad [5b]$$

therefore, the difference process is

$$D_{v,ji} = D_{vj} - D_{vi} = (\alpha_j - \alpha_i) + (\tau_v - \tau_v) + (\epsilon_{vj} - \epsilon_{vi}), \quad [5c]$$

that is,

$$D_{ji} = (\alpha_j - \alpha_i) + (\epsilon_{vj} - \epsilon_{vi}) \quad [6]$$

Equation 6 is similar to Equation 2. Importantly, it does not contain τ_v , which may be considered to be eliminated experimentally. Consequently, the subscript v is not retained in Equation 6.

It is clear that in Equation 6, as in Equation 2, $E[D_{ji}] = \alpha_j - \alpha_i$. The difference between Equations 2 and 6 is contained in the respective error terms $\epsilon'_{vj} - \epsilon'_{vi}$ and $\epsilon_{vj} - \epsilon_{vi}$. In the former, the subject parameters are retained; and the source of the correlation ρ_{ji} in the difference process of Equation 2 is attributed to the variance among subject parameters, as intimated by Thurstone. In the latter, where the subject parameters are eliminated, $\rho_{ji} = 0$; and the variance of this difference process, σ_{ji}^2 , may be expressed simply as

$$\sigma_{ji}^2 = \sigma_j^2 + \sigma_i^2 \quad [7]$$

The variance σ_{ji}^2 is unchanged whether calculated from Equation 2 or Equation 6, that is,

$$\sigma_{ji}^2 = \sigma_j'^2 + \sigma_i'^2 - 2\rho_{ji}'\sigma_j'\sigma_i' = \sigma_j^2 + \sigma_i^2. \quad [7a]$$

This situation is algebraically analogous to the methods described in elementary textbooks for finding the sampling variance of a difference between means in related or matched samples. One can either (1) involve the inter-subject variation and subtract its effect through the correlation it produces or (2) eliminate it directly through pairwise subtractions and, therefore, eliminate any correlation produced by the inter-subject variation. In the situation just cited, one can (in principle) choose which procedure to use because values corresponding to D_{vj} and D_{vi} are available. No such choice is available in the paired comparison design since (1) the pairwise subtractions are built into the design producing the outcome and (2) the outcome only indicates which of the values D_{vj} or D_{vi} is the greater. Therefore, Equation 7 rather than Equation 3 is the more appropriate general form for the variance of the difference process.

It seems surprising that Thurstone did not formalize the subject parameter whose effect he expressed so eloquently. In his analyses of paired comparison data (cf. 1927b, 1927c, 1928), however, he consistently assumed that the correlation was zero. The reformalization described above perhaps helps explain the success found with this assumption.

The correlation ρ_{ji} considered above is not the kind of correlation considered by Bock (1958) in studies of preference and choice of objects. The Bock correlation arises from a specific interaction between a subject and an item because the subject has a particular "preference," such as a taste preference, for the object. This preference for the object is then reflected in its comparisons with all other objects. However, it is less likely that subjects would have such particular preferences for some items in a test and, in any case, the assumption is open to empirical checking.

In the development of Thurstone models, it is generally assumed that the τ_v are normally distributed; this is appropriate when the intersubject variance is involved in the error variance. When it is eliminated, it is important to note that no distributional assumptions about the subject parameters need be made. This issue will be further discussed in a later section.

The Case V Specialization

Thurstone's specialization of σ_{ji}^2 for the Case V model (the relevant model in this paper) involves the assumption that $\rho_{ji} = 0$. The other specialization is that $\sigma_j^2 = \sigma_i^2$ for all i, j . In particular, let $\sigma_j^2 = \sigma_i^2 = \frac{1}{2}\sigma^2$. Then, $\sigma_{ji} = 2(\frac{1}{2}\sigma^2) = \sigma^2$. Substituting σ for σ_{ji} and changing the variable in the integrand of Equation 4 according to

$$z = \frac{t - (\alpha_j - \alpha_i)}{\sigma} \quad , \quad [8]$$

simplifies Equation 4 to

$$p\{x_{ji} = 1 | \alpha_j, \alpha_i, \sigma^2\} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}z^2} dz \cdot [9]$$

$$- (\alpha_j - \alpha_i) / \sigma$$

Another Response Function

An alternative to the normal density function is the logistic. The substitution of the logistic for the normal is commonly made in circumstances such as the above (cf. Birnbaum, 1968; Bock, 1975) because it is more tractable. In particular, the equation corresponding to Equation 9 may be readily integrated and simplified to give

$$p\{x_{ji} = 1 | \alpha_j, \alpha_i, \theta^2\} = \frac{\exp((\alpha_j - \alpha_i) / \theta)}{1 + \exp((\alpha_j - \alpha_i) / \theta)} \quad . \quad [10]$$

Furthermore, it is well known that with a constant factor adjustment according to $\sigma = 1.7\theta$, which is simply an adjustment of an arbitrary unit, the numerical evaluations of Equations 9 and 10 differ by less than .01 over the entire domain of the variable (Johnson & Kotz, 1972). With the assumption that σ and θ are constant, θ may be absorbed into the item scale values or defined arbitrarily to be unity, giving

$$p\{x_{ji} = 1 | \alpha_j, \alpha_i\} = \frac{\exp(\alpha_j - \alpha_i)}{1 + \exp(\alpha_j - \alpha_i)} \quad . \quad [11]$$

Bradley and Terry (1952) and Luce (1959) proposed the logistic model in Equation 11 for paired comparison data without reference to discriminial processes, as formulated by Thurstone.

The response functions Equations 9 and 10 have the property that the relative scale values of any pair of items of a set are invariant with respect to the values of the other items. This is an important property of the models because a given collection of items need not always be kept together for the resulting estimates of scale values to be compatible.

To summarize the development so far, one major point has shown that if the subject parameter is introduced, as in Equation 5, then it is eliminated in the paired comparison design with the consequence that (1) a rationale for the assumption that $\rho_{ji} = 0$ is provided and (2) no distributional assumptions need be made regarding the subjects. A second point was to draw attention to the development of Equation 10, an equation conceptually and numerically equivalent to Equation 9, which reflects Thurstone's Case V assumptions. Finally, it was noted that if the models of Equations 9 and 10 hold, the relative scale values of a given pair of items are independent of the scale values of other items with which they may be scaled.

The Latent Trait Formulation

In the latent trait formulation for the scaling of items, subjects respond immediately to a direct response design. The response process when subject v encounters item j may again be formalized by commencing with the discriminial process of Equation 5, that is, with

$$D_{vj} = \alpha_j - \tau_v + \varepsilon_{vj} \quad [11a]$$

However, since the response of a subject is obtained for the item without reference to any other item, it is this response process which is formalized. This may be done as follows:

1. Subject v agrees with statement j if $\tau_v > \alpha_j$, that is, $D_{vj} < 0$, and disagrees otherwise. Over hypothetical replications of responses of the same subject to the same statement, D_{vj} is assumed normally distributed with mean $\alpha_j - \tau_v$ and variance σ_j^2 .
2. σ_j^2 is assumed constant with respect to all subjects who respond to statement j .

Two points need to be brought out regarding the above formulation. First, the subject now effectively compares his/her own position on the continuum with that of the item. If he/she sees himself/herself to be above the item, he/she agrees with it; otherwise, he/she disagrees with it (cf. Lumsden, 1977, p. 478). Thus, when a subject compares his/her own position with that of an item, the process is formalized to be the same as when he/she compares the position of two items. The assumed distribution of D_{vj} is the same as that depicted in Figure 1 except that the subject parameter τ_v replaces the second item parameter α_i .

Secondly, as the latent trait formulation stands, α_j and τ_v are in reverse of what is usually seen. However, this reversal accounts for the fact that in the usual presentations, the Thurstone and latent trait scales are in reverse of each other. Accordingly, the response here is scored 1 when the subject *disagrees* rather than agrees, as is customary.

If the discrete random variable a_{vj} is defined to be 1 when $D_{vj} < 0$ (i.e., when subject v *disagrees* with statement j) and 0 otherwise, then by analogy to Equation 9,

$$p\{a_{vj} = 1 | \alpha_j, \tau_v, \sigma_j^2\} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2}z^2\right) dz \quad [12]$$

$$-\frac{(\alpha_j - \tau_v)/\sigma_j}{\sigma_j}$$

This is the normal response function extensively studied by many writers (cf. Lord, 1952, in the context of achievement test items). If the logistic response function is again substituted for the normal, then

$$p\{a_{vj} = 1 | \alpha_j, \tau_v, \theta_j\} = \frac{\exp((\alpha_j - \tau_v)/\theta_j)}{1 + \exp((\alpha_j - \tau_v)/\theta_j)}, \quad [13]$$

Applications of Equation 13 are discussed, for example, by Birnbaum (1968) and Jensema (1968).

The Rasch Formulation

In the Rasch formulation, as in Thurstone's Case V specialization, θ_j is assumed constant for all j . Therefore, as an arbitrary choice of unit, it can again be defined as unity or absorbed into the parameters. Then Equation 13 becomes

$$p\{a_{vj} = 1 | \alpha_j, \tau_v\} = \frac{\exp(\alpha_j - \tau_v)}{1 + \exp(\alpha_j - \tau_v)}. \quad [14]$$

Equation 14, which is commonly known as Rasch's simple logistic model, is identical structurally to Equation 11, but the subject parameter τ_v replaces the second item parameter α_i .

Pursuing the goal of deriving equations to estimate the scale values of the items, suppose that subject v responds to a second item i . Then

$$p\{a_{vi} = 1 | \alpha_i, \tau_v\} = \frac{\exp(\alpha_i - \tau_v)}{1 + \exp(\alpha_i - \tau_v)}. \quad [14a]$$

If a subject responds in the same way to two items, no difference between the items can be inferred. Therefore, consider the case in which the responses are different. This will occur when subject v perceives his/her position to be greater than one item but lesser than the other, that is, when he/she sees himself/herself to be between the two items. Therefore, $\{a_{vj} = 1, a_{vi} = 0\}$ or $\{a_{vj} = 0, a_{vi} = 1\}$. The respective probabilities of these pairs of responses are given by

$$p\{a_{vj} = 1, a_{vi} = 0 | \alpha_j, \alpha_i, \tau_v\} = \frac{\exp(\alpha_j - \tau_v)}{(1 + \exp(\alpha_j - \tau_v))(1 + \exp(\alpha_i - \tau_v))} \quad [15a]$$

and

$$p\{a_{vj} = 0, a_{vi} = 1 | \alpha_j, \alpha_i, \tau_v\} = \frac{\exp(\alpha_i - \tau_v)}{(1 + \exp(\alpha_j - \tau_v))(1 + \exp(\alpha_i - \tau_v))}. \quad [15b]$$

The probability of subject v perceiving his/her standing to be between two items, which can be designated by the result $a_{vj} + a_{vi} = 1$, is given by the sum of Equations 15a and 15b. That is,

$$p\{a_{vj} + a_{vi} = 1 | \alpha_j, \alpha_i, \tau_v\} = \frac{\exp(\alpha_j - \tau_v) + \exp(\alpha_i - \tau_v)}{(1 + \exp(\alpha_j - \tau_v))(1 + \exp(\alpha_i - \tau_v))}. \quad [16]$$

Suppose that subject v does perceive his/her own position to be between two items, and consider the probability that he/she perceives his/her position to be greater than that of item i (agreeing with i so

that $a_{vi} = 0$) and therefore less than that of item j (disagreeing with j so that $a_{vj} = 1$). This conditional probability is obtained by dividing Equation 16 into Equation 15a, giving, upon simplification (Rasch, 1960),

$$p\{a_{vj} = 1 | a_{vj} + a_{vi} = 1, \tau_v, \alpha_i, \tau_v\} = \frac{\exp(\alpha_j - \alpha_i)}{1 + \exp(\alpha_j - \alpha_i)} \quad [17]$$

Consider now only circumstances as above in which the subject perceives his/her position to be between two items. This is equivalent to restricting the circumstances to those from which an ordering of items can be inferred from the subject's responses. If in these restricted circumstances, the Bernoulli random variable x_{ji} is defined to take on the value 1 if the subject v perceives item j 's position to be greater than his/her position and item i 's less, then Equation 17 may be written as

$$p\{x_{ji} = 1 | \alpha_j, \alpha_i\} = \frac{\exp(\alpha_j - \alpha_i)}{1 + \exp(\alpha_j - \alpha_i)} \quad [18]$$

Note that the parameter τ , does not appear in Equation 18. It need not be retained on the left side of the equation since it has been eliminated on the right. As with the paired comparison design, the subject parameter has been eliminated. Thus, again no distributional assumptions about the subject parameters need be made. Moreover, Equation 17, the resultant pairwise estimation equation in the Rasch model, is identical to Equation 11, the estimation equation in the Thurstone Case V model when the logistic, rather than the numerically equivalent normal, function is used.

When more than two items are involved in the paired comparison design, Equation 11 can be used directly to derive either least squares (Maxwell, 1974), minimum chi-square (Bock & Jones, 1968), or maximum likelihood estimates (David, 1963). However, it cannot be generalized directly for the direct response design because the choice of overlapping subjects in different pairwise comparisons creates dependencies. Other generalizations and techniques involving least squares (Rasch, 1960) and maximum likelihood estimates (Andersen, 1973a, 1973b; Wright & Panchapakesan, 1969) have been derived. Although different algorithms need to be used in the general case, the item scale values should be equivalent in the Thurstone Case V and Rasch models.

Discussion

The above development relates the Thurstone Case V model (with the logistic function) to the Rasch model, although both models are set in wider contexts. The wider contexts were chosen to explain the relationships from first principles. The main reason for relating only the restricted models is that the Rasch model is the only latent trait one in which the subject parameters can be eliminated explicitly. The elimination of these parameters is a key issue; their explicit elimination provides for easier exposition of the main points. Further connections between Thurstone and general latent trait models are therefore deliberately not treated here.

Eliminating the Subject Parameters

The elimination of the subject parameters in the Rasch model has given rise to such expressions as "sample-free" and "population-free." In this sense, then, so is the Thurstone Case V model. Although these phrases suggest desirable properties, they are not always fully appreciated; therefore,

these terms have been avoided so far. A clarification seems in order and possible in this context.

Because subject parameters are eliminated explicitly, their *values* (individually or collectively) need not be known. As a result, no assumptions need be made about the distribution of these parameters in the population, for example, that they are normally distributed. This does not mean, of course, that the “empirical” subject population need not be specified or described or that the scale values *will* be the same in different empirical populations. In fact, it is generally vital that the empirical populations be described. But because it does not matter what the actual values of the subject parameters are in the population, the models are, in the statistical distribution sense of population, “population-free.” This is a property of the model and independent of empirical data.

It is important to note that this holds only with respect to the estimation of the *relative* scale values of the items. It does not matter if the subject parameters are large, small, or otherwise; the subjects’ impressions of the relative positions of the items should be the same. That is, items which appear more intense to one group should appear similarly more intense to the other groups. Consequently, different samples, whether drawn randomly or otherwise from the same empirical population, should give the same relative scale value estimates. In this sense the models are also “sample-free.”

However, this property of the model is true in empirical data only if the model holds in the data. Therefore, “sample-freeness” is always only a hypothesis until checked, as is the assumption that the model holds. Conversely, if the relative scale values are not sample-free, the model does not hold across these samples. This feature is often used in checks on the Rasch model, and the usual sample formation is according to total scores on the items. Therefore, the check is with respect to an interaction between relative scale values and the level of the subject parameters. This is one of the best checks since it is then known that the subject parameter values are different. However, other samples may also be formed (e.g., on the basis of sex); then the check is with respect to a sex by relative scale value interaction. If there is no interaction, then it can be inferred that the items are working in the same relative way in both groups. As a result, generalizability of the scale across the groups is demonstrated.

Note that it would be missing a crucial point to choose two *random* samples from which to check “sample-freeness” and that it would not be putting the models to any special test. Two random samples should give the same relative scale values because it is an inherent assumption of all statistical models that parameter estimates are “*random* sample-free.” For the greatest benefits, the samples should be systematically different.

There are two further important issues related to “sample-freeness” which need to be stressed. First, as implied above, such a property does not preclude actual differences in subject parameters. To give a concrete example, two sets of vocabulary items may have equivalent relative scale values with respect to groups of males and females, yet the female scores may be higher on the average than the male scores. However, such differences are essentially meaningful only if the relative scale values are equivalent in the two groups because then it is known that the items are working in the same way in the two groups. It follows, therefore, that in comparing the mean scores of two or more groups on a set of items, one should first check that the items are working in the same way across groups.

Second, the relative scale values obtained from a sample do reflect a property of that sample, namely its impressions of the relative standing of the items. In fact, this is the sense in which Thurstone described the samples and the empirical populations from which the samples were drawn. He did not concern himself with individual measurements of subjects as such. This perspective requires no change in the models, and the perspective can be employed with the Rasch model. If two samples

do not give the same relative scale values for the items, then the differences between the scale values describe differences between the samples. In comparing samples, such differences in relative scale values may be just as informative as when no differences are found.

The Wainer, Fairbank, and Hough (WFH; 1978) Study

The WFH study provides some data relevant to the equivalence of paired comparison and direct response methods. One stage of this study involved subjects from four distinct empirical community populations responding directly to a set of items describing events which require change in one's life. Fifty-one items which conformed to the Rasch model within each group and which had equivalent relative scale values across the different groups were obtained. Thus, for these respective groups, the subject parameters did not affect the relative scale values of the items; and the items worked in the same way across all groups. Furthermore, the relative scale value of any pair of items was not affected by the scale values of the other items.

WFH took advantage of this latter characteristic, selecting 5 items from the 51 and submitting them to 2 student groups of subjects in a paired comparison design. For each of these two groups, the scale values from the paired comparison data and the direct response data were equivalent after an adjustment for origin and unit. WHF were then able to adjust the remaining 46 items to the same metric. The origin and unit were different with respect to the two student groups, but these differences were taken to characterize differences between the groups. These results indicate, however, that the relative item scale values were not affected by the subject parameters in the paired comparison design.

WFH go on to say, "The advisability of using student samples to supplement a community sample may be questioned, since there are obvious status and age differences between the two. This procedure was justified by the fact that the scale . . . contained only items that scaled similarly across all the members of the community sample. . ." (p. 321). The items "scaled similarly" in the different community samples because the subject parameters were successfully eliminated through the Rasch model. But what is also critical is that the subject parameters in the different student samples were also successfully eliminated by the paired comparison design. In this connection, the WFH study shows that their 51 items are seen to have the same relative impacts by all the empirical populations they investigated (whether through the paired comparison or direct response designs) and that it does not matter what the actual subject measures were in these populations. To the degree that their populations are systematically different, the scale is generalizable; and the greater the difference in populations on demographic characteristics, the greater the generalizability demonstrated.

Conclusions

In the first chapter of their book, in a section headed "Thurstone Measurement," Bock and Jones (1968, p. 9) wrote the following:

It is a mark of the maturity of a science that the number of parameters with which it deals is small and the number of models large. In such a science, measurement can reach a high level of perfection because, first, there is generally more than one distinct model which can be said to estimate a given parametric value. This provides the possibility for independent measurement by different methods, which serves to confirm the validity of each. Second, the science may include models which make it possible to predict the effect of altered experimental conditions on the measurement procedure.

This paper, and the empirical Wainer et al. study, indicate extensions in psychological scaling of the kind indicated by Bock and Jones. Two different experimental methods, the paired comparison and the direct response, are shown to provide equivalent values of parameters specified in the respective models. Furthermore, in the replications of the experiments in the empirical study, the context varied with respect to both items and subjects, but the relative scale estimates of the specified items was invariant. Bock and Jones go on to say, "The subjects may, and usually will, be different, and the stimuli may be different. We control the variation from individual differences among subjects merely by insisting that subjects be sampled randomly from a specified population." This paper also shows, therefore, that even the requirement of *random* sampling of subjects from a population is not only unnecessarily restrictive, but with respect to some issues, not as useful as systematic sampling.

The intention of this paper was to synthesize the models used in the WFH study and to help explain some aspects of its success. That study seems critical and exemplary in the development of psychological measurement; and it is hoped that this paper will help other researchers to adapt their measurement designs, where possible, into that framework.

References

- Andersen, E.B. The numerical solution of a set of conditional estimation equations. *The Journal of the Royal Statistical Society, Series B*, 1972, 34, 42–54.
- Andersen, E.B. Conditional inference for multiple choice questionnaires. *British Journal of Mathematical and Statistical Psychology*, 1973, 26, 31–44. (a)
- Andersen, E.B. A goodness of fit test for the Rasch model. *Psychometrika*, 1973, 38, 123–140. (b)
- Birnbaum, A. Some latent trait models and their use in inferring an examinee's ability. In F.M. Lord & M.R. Novick, *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley, 1968.
- Bock, R.D. Remarks on the test of significance for paired comparisons. *Psychometrika*, 1958, 23, 323–334.
- Bock, R.D. *Multivariate statistical methods in behavioral research*. New York: McGraw-Hill, 1975.
- Bock, R.D., & Jones, L.V. *The measurement and prediction of judgement and choice*. San Francisco: Holden-Day, 1968.
- Bradley, R.A., & Terry, M.E. Rank analysis of incomplete block designs. I. The method of paired comparisons. *Biometrika*, 1952, 39, 324–345.
- David, H.A. *The method of paired comparisons*. New York: Hafner, 1963.
- Edwards, A. L. *Techniques of attitude scale construction*. New York: Appleton-Century-Crofts, 1957.
- Jensema, C.J. An application of latent trait mental test theory. *British Journal of Mathematical and Statistical Psychology*, 1974, 27, 29–48.
- Johnson, N.L., & Kotz, S. *Distributions in statistics: Continuous univariate distributions* (Vol. 3). New York: Wiley, 1972.
- Lord, F.M. A theory of test scores. *Psychometric Monograph No. 7*, 1952.
- Luce, R.D. *Individual choice behavior*. New York: Wiley, 1959.
- Lumsden, J. Person reliability. *Applied Psychological Measurement*, 1977, 1, 477–482.
- Maxwell, A.E. The logistic transformation in the analysis of paired-comparison data. *British Journal of Mathematical and Statistical Psychology*, 1974, 27, 62–71.
- Rasch, G. *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research, 1960.
- Rasch, G. On general laws and the meaning of measurement in psychology. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics*, 1961, 4, 321–334.
- Thurstone, L.L. A law of comparative judgement. *Psychological Review*, 1927, 34, 278–286. (a)
- Thurstone, L.L. The method of paired comparisons for social values. *Journal of Abnormal Social Psychology*, 1927, 21, 384–400. (b)
- Thurstone, L.L. Psychophysical analysis. *American Journal of Psychology*, 1927, 38, 368–389. (c)
- Thurstone, L.L. The measurement of opinion. *Journal of Abnormal Social Psychology*, 1928, 22, 415–430.
- Torgerson, W.S. *Theory and methods of scaling*. New York: John Wiley, 1958.
- Wainer, H., Fairbank, D.T., & Hough, R.L. Predicting the impact of simple and compound life change events. *Applied Psychological Measurement*, 1978, 2, 313–322.
- Wright, B.D. Sample-free test calibration and person measurement. *Proceedings of the 1967 Invitational Conference on Testing Problems*. Princeton, NJ: Educational Testing Service 1967, 85–101.

Wright, B.D., & Panchapakesan, N. A procedure for sample-free item analysis. *Educational and Psychological Measurement*, 1969, 29, 23–48.

Wright, B.D., & Douglas, G.A. Best procedures for sample-free item analysis. *Applied Psychological Measurement*, 1977, 1, 281–295.

Author's Address

David Andrich, Department of Education, The University of Western Australia, Nedlands, Western Australia, 6009.