Thorndike, Thurstone, and Rasch: A Comparison of Their Methods of Scaling Psychological and Educational Tests

George Engelhard, Jr.
University of Chicago and Chicago State University

The purpose of this study is to describe and compare the methods used by Thorndike, Thurstone, and Rasch for calibrating test items. Thorndike and Thurstone represent a traditional psychometric approach to this problem, whereas Rasch represents a more modern conceptualization derived from latent trait theory. These three major theorists in psychological and educational measurement were concerned with a common set of issues that seem to recur in a cyclical manner in psychometric theory. One such issue involves the invariance of item parameters. Each recognized the importance of eliminating the effects of an arbitrary sample in the estimation of item parameters. The differences generally arise from the specific methods chosen to deal with the problem. Thorndike attempted to solve the problem of item invariance by adjusting for mean differences in ability distributions. Thurstone extended Thorndike's work by proposing two adjustments which included an adjustment for differences in the dispersions of ability in addition to Thorndike's adjustment for mean differences. Rasch's method implies a third adjustment, which involves the addition of a response model for each person in the sample. Data taken from Trabue (1916) are used to illustrate and compare how Thorndike, Thurstone, and Rasch would approach a common problem, namely, the calibration of a single set of items administered to several groups.

Many great advances have been made in psychological and educational measurement within the last 20 years. Technological advances in psychometrics have made practicable many innovations that were only dreamed about in the early part of this century. Many of these recent advances in test theory and methods are described in Weiss and Davison (1981).

Although psychometricians are constantly inventing new methods, many of the actual problems and issues that are being addressed have not changed significantly since they were initially posed. In fact, many of the current solutions were adumbrated by major theorists such as Thorndike and Thurstone. An important example is the problem of item invariance.

The general problem of invariance has been succinctly stated by Wright (1968) in terms of the conditions necessary for *objective measurement*. Two of the conditions that are necessary for objective measurement are as follows:

> First, the calibration of measuring instruments must be independent of those objects that happen to be used for calibration. Second, the measurement of objects must be independent of the instrument that happens to be used for the measuring. (Wright, 1968, p. 87)

The importance of meeting these conditions, which can result in invariance, has emerged repeatedly in the history of psychological and educational measurement. The principle of invariance can be considered an essential attribute of any objective scaling method (Jones, 1960). This article is primarily

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 8, No. 1, Winter 1984, pp. 21–38

0146-6216/84/010021-18$2.05

concerned with the first condition for item invariance and with the methods used by Thorndike, Thurstone, and Rasch for meeting this condition.

Andrich (1978) has shown how Thurstone (1927a) approached the problem of invariance within the context of attitude measurement using his paired comparison method. He has shown how Thurstone's paired comparison method eliminates the sample effect experimentally, whereas Rasch's simple logistic model eliminates the sample effect statistically. The present paper extends Andrich's (1978) work by comparing how Thurstone and Rasch would approach a direct response design for the measurement of person abilities rather than attitudes.

Thurstone (1927b) believed that his method of absolute scaling met the conditions necessary for the invariance of item parameters in different groups. Thurstone illustrated his solution by reanalyzing a set of data that had been studied by Trabue (1916) using Thorndike's (1919) scaling method. Thurstone (1927b) criticized Thorndike's method of scaling because it did "not possess the one requirement of a unit of measurement, namely constancy . . . [the item difficulty scale] fluctuates from one age to another" (p. 505). Thurstone's concept of *constancy* is his version of an invariance condition and is an explicit characteristic of measurement situations that yield objective measurement. Thorndike's scale values fluctuate because the item scale values are not sample free, which violated Thurstone's (1928) insight that the "scale value of an item should be the same no matter which age group is used" (p. 119).

Although Thurstone recognized the importance of a sample-free method of item calibration that would provide invariant item parameters, his approach did not provide the final solution to the problem. Loevinger (1947) pointed out very clearly that the "subject of scaling is far from closed" (p. 43). Almost 20 years later she acknowledged that Rasch had made a major contribution to the solution of two major psychometric problems that correspond to the conditions specified by Wright (1968) for objective measurement, namely, "the ability assigned to an individual is independent of that of other members of the group and of the particular items with which he is tested; similarly for the item difficulty" (Loevinger, 1965, p. 151).

The purpose of this paper is to describe and to compare the methods used by three major measurement theorists to calibrate test items. Thorndike and Thurstone represent a traditional approach to this problem, whereas Rasch will be used to represent a more modern conceptualization derived from latent trait theory, or item response theory, as it has been called more recently (Lord, 1980). Their common concern with the elimination of arbitrary sample effects in the calibration of items serves to unite their contributions to psychological and educational measurement, as well as to highlight progress made towards the attainment of item invariance.

Description of the Three Scaling Methods

Thorndike's scaling methods, Thurstone's methods of absolute scaling, and Rasch's method of calibrating test items are described in this section. These results will then be used in the next section to highlight the similarities and differences among the three methods. The development of Thorndike's scaling method historically precedes the development of Thurstone's method of absolute scaling; however, Thurstone's methods will be described first, because Thorndike's methods can be easily derived from it.

Thurstone's Method of Absolute Scaling

The idea and motivation for developing a new method of item scaling occurred to Thurstone (1925) while he was "trying to tease out the logic of some well-known educational scales and mental age scales" (p. 433) and was finding to his dissatisfaction that the "authors of educational scale monographs seldom gave an adequate discussion of the assumptions and logic of their scale constructions" (p. 433).

In order to remedy this situation, Thurstone (1925, 1927b, 1928b) set out his ideas about item scaling in a series of major articles in the 1920s. In these articles Thurstone derived and applied his method of absolute scaling. In his words,

> It is called an absolute scaling method because it is independent of the scoring unit represented by the raw scores. The unit of measurement is the standard deviation of test ability in any given age group. This distribution is assumed, as usual, to be normal and the base line is an abstract scale of test ability independent of the raw scoring unit. (Thurstone, 1928b, p. 441)

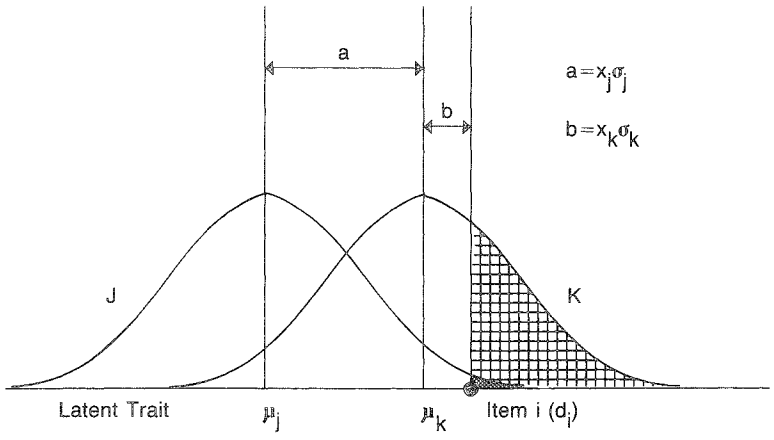
As a first step, Thurstone assumed a normal distribution of ability for any given group on the latent trait being measured. He next assumed that items could be constructed and used to define points on this scale that would define the latent trait. The location of the item was defined by the percentage of correct responses to the item for a specified group. The next step was to define a numerical value for each group with the standard deviation as the unit of measurement and the proportion of correct responses to each item within a group transformed to standard normal deviates. These standard normal deviates were called *sigma values* by Thurstone.

Once the items had been assigned a numerical value, Thurstone recognized that some adjustment must be made in order to locate each separate group with its own ability distribution on the latent trait scale. Thurstone recognized that the location of the item on the latent trait scale must be independent of the group used to calibrate the item.

In order to illustrate how Thurstone adjusted for the effects of the sample, let J represent the distribution of ability for group J and let K represent the distribution on the same latent trait for a higher ability group K. These two distributions are shown in Figure 1. Let μ_j and μ_k represent the means of the distributions in absolute scale units on the latent trait, and let σ_j and σ_k represent the standard deviations in absolute scale units. It is important to recognize the μ and σ are the unknown values in absolute scale units that must be estimated. The point on the base line (latent trait) in Figure 1 represents item i with some fixed position d_i that remains invariant regardless of the group used to define its position on the latent trait in absolute scale units. The heavily shaded area represents the proportion of people who succeeded on item i from group J, and the crosshatched plus the heavily shaded area represent the proportion of people who succeeded on item i from group K. Since group K is the more able group, a greater proportion of people in this group are successful on item i.

Now let $x_j\sigma_j$ be the distance from μ_j to item i in the standard deviation units of group J on the latent

Figure 1
Ability Distributions for Groups J and K
(Based on Figure 2 in Thurstone, 1925)



trait scale, and let $x_k\sigma_k$ be the distance between μ_k and item i in the standard deviation units of group K. It is at this point that Thurstone (1925) has an important insight:

> it is clear in [Figure 1] that $[\mu_j + x_j\sigma_j]$ must be equal to $[\mu_k + x_k\sigma_k]$ because they are measurements to the same point on the scale, both measurements representing the same test question by two different age groups. (p. 438)

Therefore,

$$\mu_j + x_{ij}\sigma_j = \mu_k + x_{ik}\sigma_k \ , \qquad [1]$$

where

μ_j = mean of group J in absolute scale units,
x_{ij} = standard normal deviate corresponding to the observed proportion of people from group J succeeding on item i,
σ_j = standard deviation of group J in absolute scale units,

and the item index i is made explicit. This can be rewritten by solving for x_{ik} for several items i, $i = 1, I$ as

$$x_{ik} = x_{ij}\,(\sigma_j / \sigma_k) + [(\mu_j - \mu_k) / \sigma_k] \ . \qquad [2]$$

Equation 2 represents the linear relationship between the observed values of x_{ik} and x_{ij}. The plot of x_{ik} against x_{ij} should be linear, and this is used by Thurstone as a check on his model. The slope of this line in absolute scale units is

$$(\sigma_j / \sigma_k) \qquad [3]$$

with an intercept of

$$[(\mu_j - \mu_k) / \sigma_k] \ . \qquad [4]$$

The observed regression of x_{ik} on x_{ij} over items is

$$x_{ik} = x_{ij}b_j + b_0 \ . \qquad [5]$$

Equation 5 can be rewritten with the correlation coefficient, r_{jk}, made explicit as follows:

$$x_{ik} = x_{ij}\,(r_{jk})\,(s_k / s_j) + [m_k - (r_{jk})\,(s_k / s_j)\,m_j] \ . \qquad [6]$$

Equation 6 becomes Equation 7 with Thurstone's assumption that the correlation coefficient is equal to one:

$$x_{ik} = x_{ij}\,(s_k / s_j) + [m_k - (s_k / s_j)\,m_j] \qquad [7]$$

where s_j = observed standard deviation of x_{ij} and m_j = observed mean of x_{ij}. The slope in observed score units is

$$(s_k / s_j) \qquad [8]$$

and the intercept is

$$[m_k - (s_k / s_j)\,m_j] \ . \qquad [9]$$

Here is another point where Thurstone has an important insight. Since Equations 2 and 7 represent the same linear relationship, one in absolute scale units and the other in observed score units, it follows that Equations 3 and 8 are identical and that Equations 4 and 9 should be equivalent. Given this assumption, the slopes can be set equal,

$$(s_k / s_j) = (\sigma_j / \sigma_k) \qquad [10]$$

or alternatively,

$$\sigma_k = \sigma_j\,(s_j / s_k) \ . \qquad [11]$$

The intercepts can also be set equal:

$$[m_k - (s_k / s_j)\,m_j] = [(\mu_j - \mu_k) / \sigma_k] \ , \qquad [12]$$

and solving for μ_k,

$$\mu_k = \sigma_j m_j - \sigma_k m_k + \mu_j \qquad [13]$$

or

$$\mu_k = \sigma_j\,[m_j - (s_j / s_k)m_k] + \mu_j \ . \qquad [14]$$

Equations 11 and 14 are Thurstone's fundamental equations for scaling items and achieving item invariance over groups. In order to solve these equations, an arbitrary origin, μ_j, and an arbitrary unit of measurement, σ_j, must be chosen. Once the origin and unit of measurement are set, the item difficulties, d_{ig}, can be estimated as follows:

$$d_{ig} = \mu_g + x_{ig}\,\sigma_g \ . \qquad [15]$$

The final weighted item difficulties, d_i, are estimated by

$$d_i = \sum_g^n w_g d_{ig} \,/\, \sum_g^n w_g \qquad [16]$$

where w_g = weighting factor based on standard error of the normal deviates and n = number of groups. Once the difficulty of each item was determined by Equation 15 in absolute scale units, Thurstone recommended taking the weighted mean of the separate item difficulties over groups, as given in Equation 16.

In summary, Thurstone's method of absolute scaling can be viewed as a two-step procedure. The first step involves the determination of the group mean and standard deviation in absolute scale units. The second step involves the estimation of the item difficulties through Equations 15 and 16.

Thorndike's Method of Scaling

Thorndike's (1919) method of scaling represents

an early attempt to transform scores on items, individuals, or "any series of facts ranked for their amounts on any traits" (p. 109) into measurements. It is important, not only as a historical landmark in mental measurement, but also as a technique which Thurstone (1927b) used as an alternative scaling method in order to document the superiority of his absolute scaling method. It is important to note that Thorndike used several methods of scaling at various points in his life. The method that is described in this section is the one used by Trabue (1916).

One of Thorndike's major contributions was his proposal for *transmuting* scores into measures on the basis of an assumed distribution (Thorndike, 1922). Once the general form of the distribution is specified, an item's position on the latent trait can be transmuted into terms of amount from a measure of central tendency in any unit of measurement based on an index of variability. Thorndike generally assumed a normal distribution of ability, and his scaling method provided a satisfactory solution for scaling problems within a single group. In fact, Thorndike's and Thurstone's methods of scaling yield essentially indentical values when applied within one group (Holzinger, 1928).

The major difference between these two scaling methods occurs in multiple group situations where a common set of items is calibrated separately in each group. Thorndike approached the problem of item invariance by adjusting for difference in the means among different groups. Thorndike's method can be expressed explicitly in a set of equations which parallel Thurstone's method of absolute scaling with the simplifying assumption that the standard deviations in each of the ability distributions are equal. Thurstone's Equation 11 takes on a value of one, and Equation 14 becomes

$$\mu_k = (m_j - m_k) + \mu_j \ . \qquad [17]$$

The item difficulty for each group, d_{ig}, can then be expressed as

$$d_{ig} = \mu_g + x_{ig} \ , \qquad [18]$$

and the final item difficulty for Thorndike's method of scaling is estimated by

$$d_i = \sum_g^n d_{ig} / n \ . \qquad [19]$$

As pointed out earlier, Thurstone's method of absolute scaling can be viewed as a two-step procedure. Thorndike's method can be viewed in a similar manner. The first step for both methods involves the determination of the group means, μ_g, and also the standard deviations, σ_g, for Thurstone's method. The second step is the determination of the item difficulties through Equations 18 and 19 for Thorndike's method of scaling.

Rasch's Method of Calibrating Test Items

During the 1950s, Rasch conducted the basic psychometric research that led to the publication in 1960 of his book *Probabilistic Models for Some Intelligence and Attainment Tests*. Rasch's method of calibrating test items was developed as an individual-centered technique as opposed to the group-based techniques used by earlier researchers, such as Thorndike and Thurstone. According to Rasch, (1960/1980)

> individual-centered statistical techniques require models in which each individual is characterized separately and from which, given adequate data, the individual parameters can be estimated. It is further essential that comparisons between individuals become independent of which particular instruments—test items or other stimuli—within the class considered have been used. Symmetrically, it ought to be possible to compare stimuli belonging to the same class—"measuring the same thing"—independent of which particular individuals within a class considered were instrumental for the comparison. (p. xx)

In traditional psychometrics, the properties of a set of items are defined in terms of the variation within specified groups. As a consequence, the properties of the items, such as the proportion correct and reliability coefficients, depend on the group being used.

Rasch's approach to measurement places a central emphasis on the concept of *specific objectivity* (Rasch, 1960/1980, 1961, 1966a, 1966b, 1977; Wright, 1968, 1977). In relation to item calibration and scaling, Rasch's aim was to develop "proba-

bilistic models in the application of which the population can be ignored'' (Rasch, 1960/1980, p. 89). His concept of specific objectivity is his version of an invariance condition and is closely related to Thurstone's concept of constancy.

The concept of specific objectivity is derived from what Rasch (1961) has termed the *principles of comparison*. Basically, these principles specify the conditions necessary for comparing individuals and for comparing items or stimuli. Rasch specified four requirements that he felt were indispensible for well-defined comparisons. The requirements are as follows:

> The comparison between two stimuli should be independent of which particular individuals were instrumental for the comparison; and it should also be independent of which stimuli within the considered class were or might also have been compared.
>
> Symmetrically, a comparison between two individuals should be independent of which particular stimuli within the class considered were instrumental for the comparison; and it should also be independent of which other individuals were also compared, on the same or on some other occasion. (Rasch, 1961, p. 331)

Rasch's measurement model for dichotomous data can be expressed as

$$\Pr\{x_{ni} = 1 \mid \beta_n, \delta_i\} = \frac{\exp(\beta_n - \delta_i)}{[1 + \exp(\beta_n - \delta_i)]} \quad [20]$$

which specifies the probability of person n with an ability of β_n, giving a correct response to item i with a difficulty of δ_i. See Rasch (1960/1980) and also Wright and Stone (1979) for a full description of the development and derivation of the model. Rasch (1966a) provides a detailed example of how this model achieves specific objectivity.

There are several methods for estimating the parameters, β_n and δ_i. The procedure used in this paper is called PROX (Wright & Stone, 1979) and was originally proposed by Cohen (1979). This estimation method was chosen because it highlights in simple closed-form equations the similarity between Rasch's approach for obtaining item invariance and sample-free item calibration, as compared to Thorndike's and Thurstone's methods.

Cohen's approximation involves making two assumptions in order to obtain equations for estimating the item difficulties and the person abilities. The first assumption is that the person ability estimates are normally distributed, and the second assumption is that the item difficulties are also approximately normally distributed. The difficulty of item i can then be expressed as

$$d_i = Y(x_i - x.) \quad [21]$$

where

x_i = logit corresponding to the observed proportion of people succeeding on item i $[\ln(1-p_i) / p_i]$,
$x.$ = mean of the item logits, and
Y = item logit expansion factor.

This can also be expressed as

$$d_i = M + x_i Y \quad [22]$$

with M set equal to $-Yx.$ and the expansion factor defined as follows:

$$Y = [(1 + V/2.89) / (1 - UV/8.35)]^{1/2} \quad [23]$$

where V = variance of observed score logits and U = variance of observed item logits.

The standard error of each item can be approximated by

$$SE(d_i) = Y[1 / (np_iq_i)]^{1/2} \quad [24]$$

where n is the number of people in the calibration sample (See Wright and Stone, 1979, for a more detailed derivation of the PROX method of estimation.) This approximation of Cohen's agrees very well with the estimates obtained using other estimation methods (Cohen, 1979; Wright & Douglas, 1977).

Comparison of Thorndike, Thurstone, and Rasch

Thorndike, Thurstone, and Rasch recognized the importance of obtaining sample-free estimates of the item parameters that would be invariant across groups. As a consequence of this, each developed a method for obtaining item invariance and for eliminating arbitrary sample effects.

The concept of an underlying latent trait plays a central role in the quest for invariance with the three scaling methods. Each method includes the

assumption that items can be located on a latent trait scale and the assumption that these items can be used to define this latent trait. The location of the item is defined by the difficulty of the item, and this difficulty is assumed to have a fixed position regardless of the group being used for the calibration.

Thorndike, Thurstone, and Rasch understood that the percentage of correct responses to an item in a particular group would not provide an invariant definition of item difficulty. Thorndike proposed a linear scale based on the transformation of the percent correct using probable errors as the unit of measurement, which he called *PE values*. Thurstone transformed the percent correct to standard normal deviates, which he called *sigma values,* and Rasch proposed using a logistic transformation of the percent correct. These transformations are very similar. In practice the PE values can be approximated very closely by dividing the standard normal deviates by a constant (.6745). This constant is derived from the relationship between the standard deviation of a normal distribution and the semi-interquartile range (Peters & Voorhis, 1940). The semi-interquartile range is used as the unit of measurement for Thorndike's PE values. In a similar fashion, the standard normal deviates can be approximated very closely by dividing the logits obtained from the logistic transformation of the percent correct by a constant (1.7).

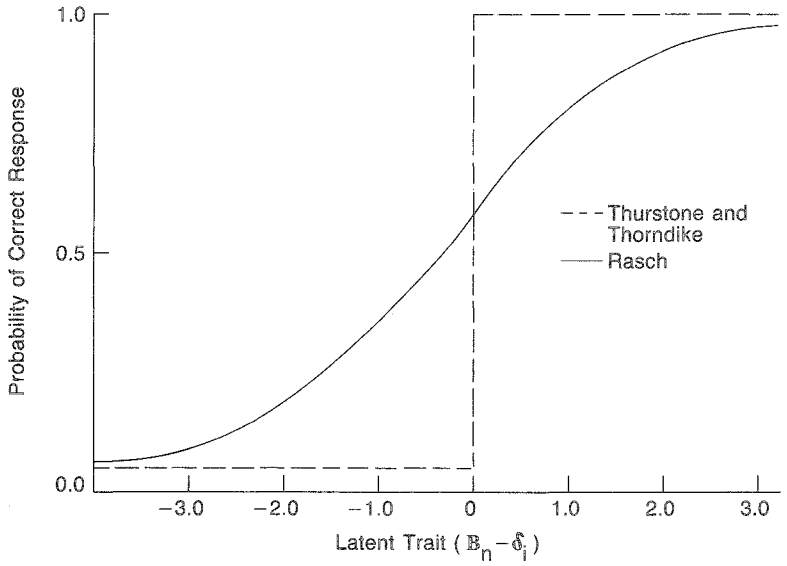
The adjustments for sample effects proposed by Thorndike, Thurstone, and Rasch are based on different levels of analysis. Thorndike and Thurstone based their adjustments on group level descriptions of the ability distributions, such as the mean and standard deviation, whereas Rasch approached the adjustment directly at the individual level of analysis. Since the adjustments are based on groups, Thorndike and Thurstone made the additional assumption that the ability distributions were normal. Given this assumption, Thorndike proceeded to use mean differences between groups to adjust for sample effects and to obtain invariant estimates of the item difficulties. Thurstone extended Thorndike's method by adding a second adjustment that allowed the standard deviations, as well as the means, to vary from group to group.

Rasch addressed the problem of item calibration directly at the individual level of analysis, and his method does not require the assumption of a normal distribution of ability. Rasch set out to develop individual-centered techniques with a "formal symmetry between items and persons" (Rasch, 1960/1980, p. 77). Rasch chose to define simultaneously the concepts of item difficulty and person ability. Rasch's method of item calibration includes an individual-level response model that can account for the probability of an individual with a certain ability succeeding on an item with a specific difficulty.

In Figure 1, the item difficulty, d_i, represents the fixed location of the item regardless of group. Thorndike's and Thurstone's methods imply that every individual to the right of the item would succeed (since their ability is greater than the item difficulty), and every individual to the left of the item would fail (since their ability is less than the item difficulty). Rasch's model describes the *probability* of obtaining a correct response on an item. This probability is defined as a function of the difference between the item difficulty and person ability. The comparison of Rasch's response model and the response model implied by Thorndike and Thurstone is given in Figure 2. Figure 2 is somewhat artificial because some people would fail due to random response error, but Thorndike and Thurstone do not propose a specific way of modeling the individual response process to take this into account.

Another difference between Thorndike and Thurstone as compared to Rasch involves the distinction between two simple questions: *Does the model fit the data?* asked by Thorndike and Thurstone and *Do the data fit the model?* asked by Rasch. One of the consequences of the way they approached the issue of model and data fit is evident in the procedures suggested for item analysis. Rasch set out the requirements necessary for specific objectivity and then advocated a very active approach to the selection and development of procedures for meeting these requirements. Essentially, Rasch suggested that items be constructed and then selected to fit his model in order to achieve objective measurement. Although Thorndike and Thurstone were certainly concerned with this issue, they placed

Figure 2
Response Models for Thorndike, Thurstone, and Rasch



a greater *relative* emphasis on the data. This emphasis on fitting models to data is the more traditional approach used in statistical analysis. Rasch advocated starting with a specific model that would lead to desirable measurement characteristics, and then creating of situations that would meet the requirements of the model. Thorndike and Thurstone placed a greater relative emphasis on the data, whereas Rasch placed a greater emphasis on his model.

Thorndike, Thurstone, and Rasch differ in the way in which they adjusted for the sample effects. Thorndike's method involves *one adjustment* for mean differences between groups in order to approximate item invariance; Thurstone proposed *two adjustments* for group differences. In Thurstone's (1927b) words,

> It is clear that in order to reduce the overlapping sentences or test items to a common base line [latent trait scale] it is necessary to make not one but two adjustments. One of these adjustments concerns the means of the several grade groups and this adjustment is made by the Thorndike scaling methods. The second adjustment which is not made by Thorndike concerns the variation in dispersion of the several groups when referred to a common scale. (p. 509)

Rasch's method implies a *third adjustment*, which involves the addition of a response model for each individual in the group. This completes the adjustment for sample effects.

The three adjustments can be seen clearly by examining Equations 15, 18, and 22. Thorndike makes an adjustment for the group mean, μ_g, in Equation 18, and Thurstone adds the additional parameter σ_g, in Equation 15, for a second adjustment based on the standard deviation of the ability distributions. Finally, Rasch adds a third adjustment, which is represented by the expansion factor, Y, in Equation 22, completing the adjustment for person ability. Other methods of estimation, such as Wright's UCON method, do not require the assumption of a normal distribution and respond in detail to the observed distributions of ability in order to approximate item invariance and sample-free item difficulties (Wright & Stone, 1979).

One final difference is Rasch's simultaneous modeling of person and item parameters. Both Thorndike and Thurstone approach the estimation

of person ability as a separate activity. Thurstone did recognize the importance of locating both individuals and items on a latent trait scale, even though he chose to treat person measurement and item calibration as separate activities. Thurstone's method of absolute scaling can, of course, be applied to the scaling of raw scores (Gulliksen, 1950), but this is distinct from the calibration of the items that led to these person scores. One of the advantages of including person measurement explicitly in the model is that it provides the opportunity to determine whether or not the *individual* being measured fits the model. This is a major advantage of all latent trait models because they provide the opportunity to define the precision of measurement for each individual (Weiss & Davison, 1981). The relationship between person measurement and item calibration is still problematic within the field of educational and psychological measurement.

The major issues that were used to compare Thorndike, Thurstone, and Rasch are summarized in Table 1.

Numerical Example

The implications of the different scaling methods proposed by Thorndike, Thurstone, and Rasch can be illustrated by examining how each would address a common problem. The problem used in this example relates to the scaling of a single set of items administered to different groups. The data used in this numerical example were taken from Trabue (1916). The items were designed to measure language ability. This data set was also used by Thurstone (1927b) to document the advantages of his method of absolute scaling.

Table 2 gives the proportion of students succeeding on Trabue's items in Grades 7 and 8. Table

Table 1
Comparison of Thorndike, Thurstone and Rasch on Major Issues

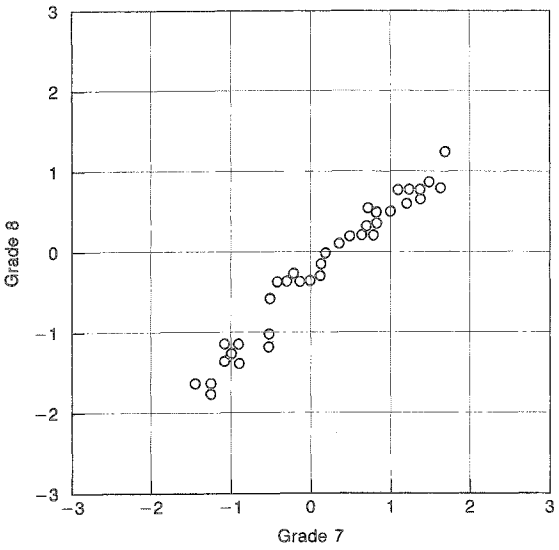
Issue	Thorndike	Thurstone	Rasch
Recognized importance of item invariance	Yes	Yes	Yes
Utilized the latent trait concept	Yes	Yes	Yes
Transformation of percent correct	PE values	Normal Deviates	Logits
Level of analysis	Group	Group	Individual
Assumed distribution of ability	Normal	Normal	None Required
Tests of fit	Model to Data	Model to Data	Data to Model
Number of adjustments	1	2	3
Item difficulties (Scale values)	$d_{ig} = \mu_g + x_{ig}$	$d_{ig} = \mu_g + x_{ig}\sigma_g$	$d_i = M + x_i Y$
Person measurement	Separate Process	Separate Process	Simultaneous Process

Table 2
Item Data for Grades 7 and 8

Items	Proportion Correct		PE Values		Standard Normal Deviates		Logits	
	7	8	7	8	7	8	7	8
3	.914	.937	-2.06	-2.36	-1.39	-1.59	-2.36	-2.70
14	.863	.883	-1.60	-1.76	-1.08	-1.19	-1.84	-2.02
15	.859	.900	-1.57	-1.92	-1.06	-1.29	-1.81	-2.20
16	.915	.941	-2.08	-2.42	-1.40	-1.63	-2.38	-2.77
18	.865	.907	-1.62	-1.99	-1.09	-1.34	-1.86	-2.28
19	.843	.903	-1.40	-1.94	-.99	-1.31	-1.68	-2.23
20	.895	.941	-1.87	-2.42	-1.26	-1.63	-2.14	-2.77
21	.837	.887	-1.43	-1.71	-.96	-1.21	-1.64	-2.06
23	.741	.850	-.92	-1.51	-.62	-1.02	-1.05	-1.73
24	.751	.882	-.96	-1.75	-.65	-1.18	-1.10	-2.01
25	.688	.738	-.69	-.90	-.46	-.61	-.79	-1.04
26	.610	.661	-.39	-.58	-.26	-.39	-.45	-.67
27	,567	.680	-.24	-.66	-.16	-.44	-.27	-.75
28	.550	.668	-.18	-.69	-.12	-.41	-.20	-.70
30	.548	.673	-.17	-.63	-.11	-.42	-.19	-.72
31	.540	.641	-.14	-.50	-.09	-.34	-.16	-.58
32	.131	.211	1.65	1.15	1.11	.78	1.89	1.32
33	.328	.441	.62	.21	.42	.14	.72	.24
34	.459	.620	.14	-.43	.10	-.29	.16	-.49
35	.450	.572	.18	-.25	.12	-.17	.20	-.29
36	.220	.370	1.10	.46	.74	.31	1.26	.53
37	.425	.501	.26	.00	.18	.00	.30	.00
38	.303	.384	.73	.41	.49	.28	.83	.47
39	.279	.422	.83	.27	.56	.18	.95	.31
40	.238	.389	1.01	.39	.68	.27	1.16	.45
41	.219	.340	1.11	.58	.75	.39	1.27	.66
42	.220	.288	1.09	.79	.74	.53	1.25	.90
43	.228	.395	1.06	.37	.72	.25	1.22	.43
44	.145	.293	1.55	.77	1.04	.52	1.77	.88
45	.125	.261	1.70	.91	1.14	.61	1.95	1.04
46	.103	.203	1.89	1.19	1.27	.80	2.16	1.37
47	.114	.217	1.79	1.12	1.21	.75	2.05	1.28
48	.077	.203	2.16	1.19	1.46	.80	2.48	1.37
49	.084	.178	2.08	1.33	1.40	.90	2.39	1.53
50	.103	.215	1.89	1.13	1.27	.76	2.16	1.30
51	.059	.104	2.42	1.88	1.63	1.27	2.77	2.15
Mean			.22	-.28	.15	-.19	.25	-.33
S.D.			1.37	1.26	.92	.85	1.57	1.44

2 also gives the transformations of these percents to Thorndike's PE values, Thurstone's standard normal deviates, and Rasch's logits. These transformations have a very simple relationship to one another, as was pointed out earlier. The proportion correct, p_i, can be converted to logits, L_i, using the

Figure 3
Relationship Between Thurstone's Scaling Values Estimated Separately in Grades 7 and 8 (Based on Figure 3 in Thurstone, 1927b)



following equation:

$$L_i = \ln[(1 - p_i)/p_i] \quad . \qquad [25]$$

The logits can be used to approximate standard normal deviates, N_i, by using Equation 26.

$$N_i = L_i / 1.7 \qquad [26]$$

The PE values, PE_i, can be approximated from the standard normal deviates as follows:

$$PE_i = N_i / .6745 \quad . \qquad [27]$$

The differences between these transformations are trivial in practical measurement situations, except for the computational advantages of the logits. The approximations given above agree quite closely with the actual transformations used by Thorndike and Thurstone. The approximations highlight the similarity among the three methods.

Both Thorndike and Thurstone would calibrate the items separately for each group and then adjust for group differences in order to obtain sample invariant item difficulties. Since their methods are very similar, and in fact can be shown to yield essentially identical values in this two-group situation, only Thurstone's method will be discussed here.

Thurstone's next step, after converting the proportions correct to standard normal deviates, would be to plot these values for adjacent grade groups. The relationship between the standard normal deviates calculated separately in each grade is given in Figure 3. Since the relationship is clearly linear, Thurstone would then proceed to use Equation 14. Table 2 gives the basic data necessary for solving Equation 14. The mean in absolute scale units for Grade 8 is 6.09. This value is obtained by setting the mean and standard deviation in absolute scale units for Grade 7 to 5.66 and 1.22, respectively $\{\mu_8 = 1.22[.15 - (.92/.85)(-.19)] + 5.66\}$. The standard deviation in absolute scale units for Grade 8 can be estimated by Equation 11 and is equal to 1.32 $\{\sigma_8 = 1.22(.92/.85)\}$.

The first step in Thurstone's method of absolute scaling is completed, namely, the estimation of the mean and standard deviation in absolute scale units. The second step involves solving Equation 15 for each item. For example, the difficulty or scale value for Item 14 in Grade 8 is equal to 4.52 $\{d_{14,8} = 6.09 + 1.32(-1.19)\}$. Thurstone would then compute the item scale values separately in each grade group and then average these values to obtain the final item difficulty as shown in Equation 16.

Rasch, by contrast with Thurstone, would approach the problem of estimating item difficulties in a multiple-group situation in a different manner. Rasch (1960/1980) felt that the "best estimate of degree of difficulty of an item is found by using the frequency of correct answers in the whole population, irrespective of which persons solved the item correctly" (p. 77). The data necessary for illustrating Rasch's approach to item calibration are given in Table 3.

In order to make Rasch's approach to item calibration clearer, the data on Item 3 will be used. The expansion factor, Y, is equal to 1.059: $Y = \{(1 + .189/2.89)/[1 - (.189)(2.19)/8.35]\}^{1/2}$, where the observed variance of person scores in logits, V, is equal to .189, and the variance of the observed item logits, U, is equal to 2.19. The observed variance of person scores, V, is a rough approximation based on slightly different data, because a complete item by person matrix was not available in Trabue (1916). The difficulty for Item 3 can be estimated using Equation 22 and is equal to -2.65 $\{d_3 = (-2.50)(1.059)\}$, where M is set equal to zero. The standard error can be estimated by Equation

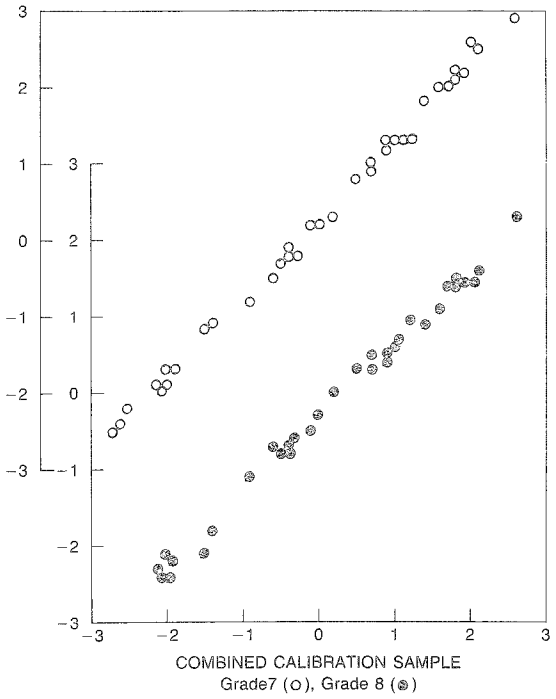
Table 3
Item Data for Rasch's Method of Scaling

Item	Item Difficulties		
	7	8	Combined
3	-2.44	-2.86	-2.65
14	-1.91	-2.14	-2.02
15	-1.88	-2.33	-2.08
16	-2.47	-2.93	-2.68
18	-1.93	-2.41	-2.05
19	-1.74	-2.36	-1.99
20	-2.22	-2.93	-2.52
21	-1.70	-2.18	-1.91
23	-1.09	-1.83	-1.39
24	-1.14	-2.13	-1.50
25	-.82	-1.10	-.94
26	-.47	-.71	-.57
27	-.28	-.79	-.49
28	-.21	-.74	-.42
30	-.20	-.76	-.43
31	-.17	-.61	-.35
32	1.96	1.40	1.73
33	.75	.25	.54
34	.17	-.52	-.11
35	.21	-.31	.00
36	1.31	.56	.98
37	.31	.00	.18
38	.86	.50	.72
39	.98	.33	.71
40	1.20	.48	.89
41	1.32	.70	1.05
42	1.30	.95	1.16
43	1.26	.45	.91
44	1.83	.93	1.42
45	2.02	1.10	1.59
46	2.24	1.45	1.87
47	2.12	1.35	1.78
48	2.57	1.45	2.01
49	2.48	1.62	2.08
50	2.24	1.38	1.83
51	2.87	2.27	2.63
Mean	.26	-.35	.00
S.D.	1.63	1.52	1.57
N	1456	1427	2883

24 and is equal to .074 $\{SE(d_3) = 1.059[1/(2883)(.924)(.076)]^{1/2}\}$.

Although a complete response matrix was not available for Trabue's data, it is still possible to examine the fit of the data to Rasch's model. The empirical test of the model is based on the main characteristic of his model, namely, if sample-free calibration is achieved, then the relative values of

Figure 4
Rasch's Graphic Control of His Model



the item difficulties should remain invariant across groups. In order to test this hypothesis, Rasch would plot the item difficulties obtained in separate groups on a reference calibration based on the item difficulties obtained from the combined groups. If the hypothesis of sample-free item calibration is supported, then the points should define parallel lines with slopes of one. There are, of course, better methods of examining the fit of the data than this graphic approach, (Andersen, 1973; Wright & Stone, 1979) but this method will suffice for the purposes of this example.

The data for examining the model using what Rasch (1960/1980) termed *control of the model* is given in Table 3, and the plot of these data is given in Figure 4. The items do seem to be invariant according to the conditions specified by Rasch for specific objectivity. This lends support to the contention that sample invariant item difficulties have been obtained.

The final difficulties obtained by Thorndike, Thurstone, and Rasch are given in Table 4. These estimates are highly consistent. The plots showing the high degree of consistency among the three different methods are given in Figures 5, 6, and 7. The correlation between item difficulties obtained by Thorndike and those obtained by Thurstone is .987, and the correlation between the item difficulties obtained by Thorndike and by Rasch is .986. The correlation between the item difficulties obtained by Thurstone and Rasch is .999. It is clear that the scaling methods of Thorndike, Thurstone, and Rasch show a high degree of agreement and that except for arbitrary differences in location and scale the results are very similar.

Discussion and Summary

Great progress has been made towards the solution of a number of significant problems in psychological and educational measurement. Occasionally, it is useful to look at some of the earlier work on these problems and to review the solutions proposed by earlier measurement theorists. One purpose of this paper was to provide an historical review of the work of major measurement theorists and to highlight their contributions to the solution of problems related to item invariance and sample-free item calibration. Thorndike and Thurstone were chosen to represent earlier attempts at adjusting for arbitrary sample effects on the calibration of item difficulties. Rasch was chosen to represent the more recent contributions of latent trait theory to the solution of this problem.

An important aspect of each of the three scaling methods was the provision of an adjustment for arbitrary sample effects. The goal of item invariance was characteristic of each method. They were also very similar in that the concept of a latent trait played a significant role in the methods for approximating sample-free item difficulties. Thorndike, Thurstone, and Rasch chose different ways of transforming the proportion correct to linear scales, but these differences are trivial in practical measurement situations.

The first significant difference between Thorndike and Thurstone versus Rasch emerges in the choice of a level of analysis. Thorndike and Thurstone used a group level of analysis, whereas Rasch chose to develop a model at the individual level of

Table 4
Final Item Difficulties

Item	Thorndike	Thurstone	Rasch
2	1.38	2.898	-2.40
3	3.33	3.579	-1.32
5	2.52	3.384	-1.56
6	1.98	3.059	-2.01
7	3.34	3.752	-1.05
8	2.94	3.591	-1.26
9	3.76	3.949	-.77
10	3.41	3.806	-.96
11	3.31	3.775	-1.01
12	3.58	3.846	-.92
13	4.03	4.114	-.57
14	4.88	4.135	-.55
15	4.81	4.396	-.19
16	4.15	4.133	-.53
17	3.66	3.908	-.83
18	4.42	4.329	-.28
19	4.47	4.055	-.65
20	4.09	4.094	-.58
21	4.04	4.307	-.31
22	4.26	4.362	-.25
23	5.40	4.797	.33
24	5.69	4.840	.38
25	6.32	4.852	.35
26	7.00	5.055	.63
27	6.67	5.181	.83
28	7.04	5.441	1.14
29	4.12	4.212	-.43
30	6.95	5.434	1.14
31	6.50	5.212	.89
33	8.38	6.335	2.34
34	7.31	5.761	1.58
35	7.42	5.806	1.64
37	7.85	6.012	1.92
38	8.42	6.470	2.58
39	8.58	6.537	2.66
Mean	5.04	4.556	.00
S.D.	1.94	.960	1.28

analysis. This decision had an impact on the subsequent steps taken by each in their methods of scaling. Thorndike and Thurstone both assumed that the ability distributions were normal; Rasch's method does not require any distributional assumptions.

Another difference between Thorndike and Thurstone versus Rasch was the relative emphasis placed on the role of data and models in the tests of fit. Thorndike and Thurstone used a more traditional statistical approach to data analysis and therefore placed a greater relative emphasis on the

Figure 5
Comparison of Thorndike's and Thurstone's Final Item Difficulties

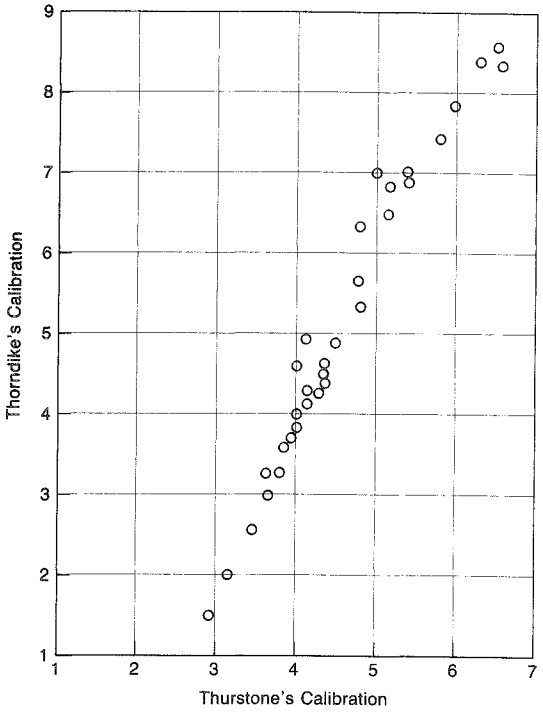
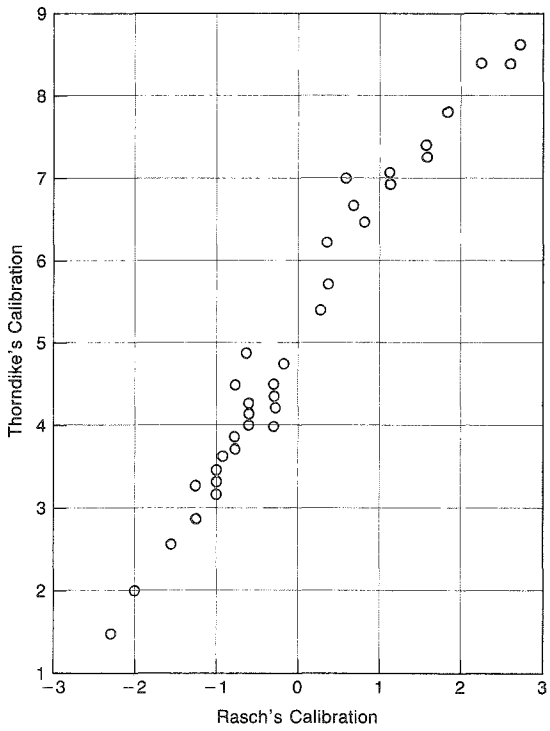


Figure 6
Comparison of Thorndike's and Rasch's Final Item Difficulties

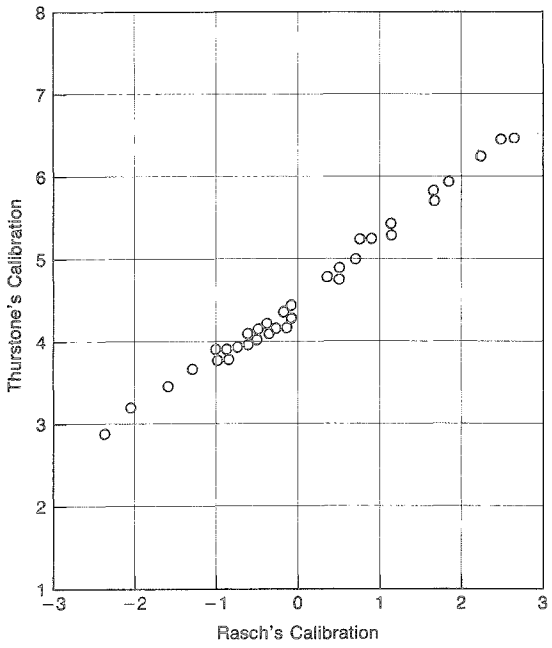


importance of fitting models to data. Rasch, by contrast, approached the issue of fit by beginning with a psychometric model that possessed what Rasch determined were desirable measurement characteristics. Rasch placed a greater relative emphasis on his model rather than on the data. Later in his career Thurstone (1947) also became concerned with the fit of data to his model and suggested that misfitting or deviate items be ignored. Misfitting items were defined by examining data plotted like Figure 3 and by ignoring any items that were "noticeably off the curve" (p. 104). This approach is essentially equivalent to Rasch's method for controlling his model.

Thorndike, Thurstone, and Rasch differ in the number of adjustments that they advocated to approximate item invariance. Thorndike made one adjustment for differences in the means between two ability distributions, whereas Thurstone made two adjustments. Thurstone adjusted for differences in the standard deviations, as well as the means. Rasch's method of item calibration implies a third adjustment, not made by Thurstone, concerning the variation in an individual's response that requires the introduction of a response model. The implementation of these adjustments can be seen in the equations used for calibrating the items, which are summarized in Table 1.

A final difference between the three methods concerns the approach suggested for person measurement. Thorndike and Thurstone chose to model the person measurement process separately from the calibration of items. Rasch chose to model simultaneously the person and item parameters that govern the probability of a person succeeding on an item. The implications of this choice probably define the most significant differences between the methods. For example, the prime advantage of including the individual directly in an item response model is that it provides the opportunity to examine

Figure 7
Comparison of Thurstone's and Rasch's Final Item Difficulties



whether or not the person's responses are congruent with the model.

In the numerical example it was shown that when Thorndike's, Thurstone's, and Rasch's methods are applied to a common data set, the results are very similar. This similarity is at least partly due to the well-behaved nature of Trabue's (1916) data set, as well as to the similarities between the scaling models. The well-behaved nature of this data set is indicated by Figure 3, which would be used by Thurstone to examine the fit of his model to the data. It is also indicated by Figure 4, which would be used by Rasch to examine the fit of the data to his model. The methods would probably yield different estimates of item difficulties under less desirable measurement conditions. It is reassuring that the methods do converge on essentially equivalent results in this particular case. It would be interesting to see how the methods compare under less favorable measurement conditions, where, for example, the assumptions about normality do not hold.

In an earlier comparison of Thorndike's and Thurstone's methods of scaling items, Loevinger (1947) concluded that "despite the similarity of the scaling methods of Thorndike and Thurstone, they have very different conceptions of what they are doing" (p. 23). The evidence presented in this paper suggest that this is not really the case. In fact, when Thorndike and Thurstone are compared to Rasch, the great similarities between their measurement philosophies and even their methods become evident. It seems that Thorndike, Thurstone, and Rasch had very similar measurement philosophies and many similar goals and that the major differences among them involve the specific methods chosen to implement their measurement models. It also seems that Thorndike and Thurstone were on the right track in their pursuit of item invariance and their contributions provided a framework for other advances in psychological and educational measurement.

Andrich (1978) attributed the correspondence between Thurstone's method of paired comparison and Rasch's method to the fact that Thurstone's method eliminates *experimentally* the effects of the sample, whereas Rasch's method eliminates the effects of the sample *statistically*. The present paper includes a comparison of how Thurstone and Rasch would both approach a direct response design. The agreement between Thurstone's and Rasch's methods for this type of design is also quite high. Thorndike's one adjustment and Thurstone's two adjustments to obtain item invariance represent attempts to eliminate *statistically* the sample effects. Thorndike, Thurstone, and Rasch each provide statistical adjustments for the arbitrary effects of the sample. Since the data are well behaved, the adjustments of Thorndike, Thurstone, and Rasch provide an adequate approximation to item invariance.

Rasch's method of item calibration provides a more complete adjustment for the sample distribution of ability, which includes the introduction of a response model to account for the variation in an individual's responses that can sustain what Wright (1968) has termed *sample-free item calibration*. Rasch's method can be viewed as similar to Thorndike's scaling method and Thurstone's

method of absolute scaling with a response model included and the adjustment for the arbitrary effects of the sample completed.

In general, the similarity between the three methods is a positive indicator. It shows how the Rasch model and other latent trait models fit into the history of psychological and educational measurement. The issue of item invariance and sample-free estimation is still important, and one that may have not been totally resolved. Wood (1976) has pointed out that although

> much play is made of the 'invariance' of item parameter estimates in latent trait models supportive evidence is conspicuous by its absence. In fact, this is one of the greyest areas of test theory. (p. 252)

The work of Thorndike, Thurstone, and Rasch represents some of the major milestones in the quest for item invariance. It seems that further work on this issue may still be required.

References

Andersen, E. B. (1973). A goodness of fit test for the Rasch model. *Psychometrika, 38,* 123–140.

Andrich, D. (1978). Relationships between the Thurstone and Rasch approaches to item scaling. *Applied Psychological Measurement, 2,* 451–462.

Cohen, L. (1979). Approximate expressions for parameter estimates in the Rasch model. *British Journal of Mathematical and Statistical Psychology, 32,* 113–120.

Gulliksen, H. (1950). *Theory of mental tests*. New York: John Wiley & Sons, Inc.

Holzinger, K. J. (1928). I. Some comments on Professor Thurstone's method of determining the scale values of tests items. *Journal of Educational Psychology, 19,* 112–117.

Jones, L. V. (1960). Some invariant findings under the method of successive intervals. In H. Gulliksen & S. Messick (Eds.), *Psychological scaling: Theory and applications*. New York: John Wiley & Sons, Inc.

Loevinger, J. (1947). A systematic approach to the construction and evaluation of tests of ability. *Psychological Monographs, 61,* No. 4.

Loevinger, J. (1965). Person and population as psychometric concepts. *Psychological Review, 72,* 143–155.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale NJ: Erlbaum Assoc.

Peters, C. C., & Voorhis, W. R. (1940). *Statistical procedures and their mathematical bases*. New York: McGraw-Hill Inc.

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

Rasch, G. (1961). On general laws and the meaning of measurement in psychology. In J. Neyman (Ed.), *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*. Berkeley: University of California Press.

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

Rasch, G. (1966b). An item analysis which takes individual differences into account. *British Journal of Mathematical and Statistical Psychology, 19,* Part 1, 49–57.

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

Thorndike, E. L. (1919). *An introduction to the theory of mental and social measurements*. New York: Columbia University, Teachers College.

Thorndike, E. L. (1922). On finding equivalent scores in tests of intelligence. *Journal of Applied Psychology, 6,* 29–33.

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

Thurstone, L. L. (1927a). The method of paired comparisons for social values. *Journal of Abnormal and Social Psychology, 21,* 384–400.

Thurstone, L. L. (1927b). The unit of measurement in educational scales. *Journal of Educational Psychology, 18,* 505–524.

Thurstone, L. L. (1928a). II. Comment by L. L. Thurstone. *Journal of Educational Psychology, 19,* 117–124.

Thurstone, L. L. (1928b). Scale construction with weighted observations. *Journal of Educational Psychology, 19,* 441–453.

Thurstone, L. L. (1947). The calibration of items. *American Psychologist, 2,* 103–104.

Trabue, M. R. (1916). Completion-test language scales. *Contributions to Education, No. 77*. New York: Columbia University, Teachers College.

Weiss, D. J., & Davison, M. L. (1981). Test theory and methods, *Annual Review of Psychology, 32,* 629–658.

Wood, R. (1976). Trait measurement and item banks. In D.N.M. de Gruijter & L.J.T. van der Kamp (Eds.), *Advances in Psychological and Educational Measurement*. New York: John Wiley & Sons.

Wright, B. D. (1968). Sample-free test calibration and

person measurement. In *Proceedings of the 1967 Invitational Conference on Testing Problems*. Princeton NJ: Educational Testing Service.

Wright, B. D. (1977). Solving measurement problems with the Rasch model. *Journal of Educational Measurement, 14*, 97–116.

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

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

Acknowledgments

The author is grateful to Benjamin D. Wright for suggesting that the Rasch model represents a "third adjustment." Judith A. Monsaas provided many helpful comments.

Author's Address

Send requests for reprints or further information to George Engelhard, Jr., Office of Institutional Research and Evaluation, Chicago State University, Ninety-Fifth at King Drive, Chicago IL 60628, U.S.A.