JOURNAL OF EDUCATIONAL MEASUREMENT
VOL. 11, NO. 2 FALL 1974

# THE NATURE OF OBJECTIVITY WITH THE RASCH MODEL

SUSAN E. WHITELY[1] AND RENÉ V. DAWIS[2]
*University of Minnesota*

Although it has been claimed that the Rasch model leads to a higher degree of objectivity in measurement than has been previously possible, this model has had little impact on test development. Population-invariant item and ability calibrations, together with the statistical equivalency of any two item subsets, are supposedly possible if the item pool has been calibrated by the Rasch model. Initial research has been encouraging, but the implications of underlying assumptions and operational computations in the Rasch model for trait theory have not been clear from previous work. The current paper presents an analysis of the conditions under which the claims of objectivity will be substantiated, with special emphasis on the nature of equivalent forms. It is concluded that the real advantages of the Rasch model will not be apparent until the technology of trait measurement becomes more sophisticated.

A procedure for a new kind of item analysis, based on the Rasch (1960, 1961, 1966a, 1966b) logistic model, is now available for use in developing measures of unidimensional traits. Wright (1967), one of the first researchers to operationalize the Rasch model, claims that the use of this model leads to an objectivity in measurement which is not possible under classical approaches to test development. The objectivity, according to Rasch (1961), results from two basic features of the model: 1) the calibration of the test items is independent of the sample and 2) the measurement of a person on the latent trait is independent of the particular items used.

A psychological test having these general characteristics would become directly analogous to a yardstick that measures the length of objects. That is, the intervals on the yardstick are independent of the length of the objects and the length of individual objects is interpretable without respect to which particular yardstick is used. In contrast, tests developed according to the classical model have neither characteristic. The score obtained by a person is not interpretable without referring to both some norm group and the particular test form used.

Wright and Panchapakesan (1969) claim that objective measurement is now possible because their estimation techniques for the Rasch parameters yield tests with the following specific properties: 1) the estimates of the item difficulty parameter will not vary significantly over different samples of people, 2) the estimates of a person's ability, given a certain raw score, will be invariant over different samples, and 3) estimates of a person's ability from any calibrated subset of items will be statistically equivalent. If these properties are truly now possible through application of the Rasch model, it would seem that mental measurement would be revolutionized. No longer would equivalent forms need to be carefully developed, since measurement is instrument independent and any two subsets of the calibrated item pool could be used as alternative

[1]Now at the Psychology Department, University of Kansas, Lawrence, Kansas.

[2]This study was supported by an Office of Naval Research contract to Rene V. Dawis through the Center for the Study of Organizational Performance and Human Effectiveness at the University of Minnesota.

instruments. Similarly, independence of measurement from a particular population distribution implies that tests can be used for persons dissimilar from the standardization population without the necessity of collecting new norms.

To date, however, the Rasch model has had little apparent impact on test development. The reasons for this are not clear, particularly since initial research has been encouraging. Both item and ability parameters have been found to be invariant over different nonrandom samples (Anderson, Kearney, & Everett, 1968; Brooks, 1965; Tinsley, 1971). Furthermore, the model appears to be robust with respect to several of the underlying assumptions (Panchapakesan, 1969). However, little evidence on the equivalency of item subsets has been presented, nor is it clear from Wright and Panchapakesan's (1969) paper how application of the Rasch model yields either item-invariance or sample-invariance of the estimated parameters.

The major purpose of the present paper is to determine how the properties of the Rasch model, estimation procedures, and trait data interact to produce item- and sample-invariant parameters. The equivalency of item subsets will be given special attention by presenting some empirical data in addition to determining thoroughly the nature of subset equivalency.

## THE RASCH MODEL

The Rasch model is a latent structure model which is based on the outcome of the encounter between persons and items. The model seeks to reproduce, as accurately as possible, the probabilities of passing items, in the cells of an item by score group matrix, where persons obtaining the same raw score are grouped together. Table 1 presents an item by score group matrix in which $k$ items are ordered by their difficulty level and $k - 1$ score groups by obtained raw scores. The score groups for which all items are either passed or failed are excluded from the matrix, since these extreme score groups provide no differential information about the items. The cell entries represent the probability, $P_{ij}$, that item $i$ will be passed by score group $j$. The Rasch model is a function which is designed to reproduce these proportions of probabilities by use of only two parameters, item easiness and person ability, in the following manner:

$$P_{ij} = \frac{A_j \times E_i}{1 + A_j \times E_i} \tag{1}$$

where $A_j$ = ability parameter for score group $j$ and $E_i$ = easiness parameter for item $i$. These parameters, $A_j$ and $E_i$, are latent marginals associated with the individual score groups and items, respectively.

Scaling a test through the use of the Rasch model may be contrasted to traditional test development techniques. The traditional indices associated with persons and items, number correct and percent passing respectively, are direct linear calculations from the data. In contrast, the task in calibration with the Rasch model is to estimate the person and item parameters so that the probabilities of each person (score group) passing each item will be accurately reproduced. Although the Rasch ability parameters are monotonically related to raw scores, a scoring table must be used to convert one index to the other. Once an item pool is calibrated, the scoring table may be used to determine the Rasch ability equivalents of new raw scores. The Rasch ability parameters may also be determined for raw scores from tests which use only some

Table 1

Item by Score Group Probability Matrix

| | | Score Level (Raw Score Groups) | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | | j = 1, k-1 |
| | | 1 | 2 | 3 | ... | 1 | ... | m | ... | k-1 |
| i = 1,k | 1 | $P_{11}$ | $P_{12}$ | $P_{13}$ | ... | $P_{11}$ | ... | $P_{1m}$ | ... | $P_{1(k-1)}$ |
| | 2 | $P_{21}$ | $P_{22}$ | $P_{23}$ | ... | $P_{21}$ | ... | $P_{2m}$ | ... | $P_{2(k-1)}$ |
| | . | . | . | . | | . | | . | | . |
| | . | . | . | . | | . | | . | | . |
| | . | . | . | . | | . | | . | | . |
| Items | g | $P_{g1}$ | $P_{g2}$ | $P_{g3}$ | | $P_{g1}$ | | $P_{gm}$ | | $P_{g(k-1)}$ |
| | . | . | . | . | | . | | . | | . |
| | . | . | . | . | | . | | . | | . |
| | . | . | . | . | | . | | . | | . |
| | k | $P_{k1}$ | $P_{k2}$ | $P_{k3}$ | | $P_{k1}$ | | $P_{km}$ | | $P_{k(k-1)}$ |

subset of the calibrated items. Separate scoring tables can be produced for any subset of items.

*Assumptions and Implications*

Rasch (1961) makes two basic assumptions in recommending the application of his model to psychological test data: 1) the model, equation (1), characterizes the data and 2) subjects and items are locally independent. The first assumption has several implications, to be described in some detail, while the second assumption concerns the experimental conditions of the test situation. Independence of subjects means that the item responses of any given person do not affect the responses of any other person. Independence of items, on the other hand, means that a person's responses to preceding items do not affect his responses to later items. Thus, the probabilities a person will pass the various individual items must remain invariant, regardless if the ability test contains the whole item pool or only some subset of items.

The first assumption, equation (1) is true, has several implications for the kind of data to which the Rasch model may be appropriately applied. The most basic implication is unidimensionality of the item pool. This means that if subjects are grouped according to raw score, within each group, there will be no remaining significant correlations between items. Thus, all of the covariation between the items is accounted for by variation of persons on the latent trait to be measured.

Referring again to Table 1, the item by score group matrix, unidimensionality implies that for each item, $P_{i1}$ is less than $P_{i2}$ and $P_{i2}$ is less than $P_{i3}$ and so on to $P_{i,k-1}$, so that the probability of passing the item increases regularly with total score. Each item, then orders subjects in the same way.

Another implication required for conjoint measurement of subjects and items, is that items are ordered in the same way within each score group. On the item by score group matrix, this implies that $P_{1j}$ is less than $P_{2j}$ and $P_{2j}$ is less than $P_{3j}$, etc. to $P_{kj}$, within each score group. A further implication is that probability of passing items must increase over score groups in accordance with the item characteristics curve specified by the Rasch logistic model. The shape of this curve closely approximates the cumulative normal distribution curve.

Test data for which the Rasch model is applicable must have two other properties. First, all items must have equal discrimination, that is, the rate at which the probability of passing the item increases with total score must be equal for all items. The Rasch model does not contain a parameter for item discrimination. Second, there must be minimal guessing so that the probability of passing an item by chance is minimized.

As summarized by Wright and Panchapakesan (1969), use of the Rasch model implies that the only way in which items differ is easiness. Although on the surface this seems to lead to a very restricted applicability of the model, several researchers have claimed the model is robust with respect to significant departures from these implied properties (Anderson, Kearney, & Everett, 1968; Panchapakesan, 1969; Wright & Panchapakesan, 1969). However, as will be pointed out in this paper, the degree to which calibrations by the Rasch model will yield the objective measurement characteristics claimed by Wright and Panchapakesan (1969), depends directly on how much the data departs from some of these implied properties and the assumption of local independence. The next few sections will be devoted to explaining the model and estimation techniques.

*Estimating the Parameters*

An understanding of how the item and person parameters are determined from the Rasch model necessitates converting the cell probabilities, as on Table 1, into likelihood ratios. Likelihood ratios are simply betting odds, the ratio of the probability of passing to the probability of failing. In Wright and Panchapakesan's (1969) estimation procedure for Rasch item and person parameters, a correction factor is added to the cell likelihoods to prevent infinite values from occurring when all members of the score group pass the item.[3] For conceptual clarity, only the uncorrected cell likelihoods, without regard to estimation difficulties, will be used in this discussion. Given $P_{ij}$, the

[3]The cell likelihood values are corrected by the relative frequency of the score group, as given by the following equation:

$$L = \frac{a_{ij} + w}{r_{ij} - a_{ij} + w},$$

where $L$ = corrected cell likelihood $a_{ij}$ = number of persons in score group $j$ passing item $i$, $r_{ij} - a_{ij}$ = number of persons in score group $j$ failing item $i$, and $w$ = percentage of total calibrating sample obtaining score $j$.

cell likelihoods for a given item and a given score group are reproduced by the simple product of item easiness and person ability values as follows:

$$\frac{P_{ij}}{1 - P_{ij}} = A_j \times E_i. \tag{2}$$

Accordingly, the likelihoods in the cells of the item by score group matrix are reproduced from the values associated with the row and column marginals. The person ability value represents an indication of the likelihood that a person will pass any item in the set, whereas the item easiness value is an indication of the likelihood the item will be passed by persons in any score group. This will become more obvious by a brief consideration of the computations of the parameters.

The values for the Rasch model person and item parameters are usually reported as log likelihoods, rather than simple likelihoods. Similarly, the cell likelihoods are also converted to logarithms. The alternate set of values, $t_{ij}$, $b_j$, and $d_i$, are derived from the parameters defined in equation (2) as follows:

$$t_{ij} = \log \frac{P_{ij}}{1 - P_{ij}},$$

$$b_j = \log (A_j), \quad \text{and}$$

$$d_i = \log (E_i).$$

Using log likelihoods rather than simple likelihoods has two advantages. The first is the obvious computational advantage. Second, the estimate of the log likelihood of any cell in the matrix, $t_{ij}$, is the simple sum of $\log A_j$ and $\log E_1$ as follows:

$$t_{ij} = b_j + d_i. \tag{3}$$

Thus, on the logarithmic scale, the likelihood that a person will pass an item is given by the simple addition of his ability and the item's easiness.[4]

The initial estimates for items, $d_i$, are obtained directly from averaging the log likelihoods of passing the items over score groups, and then subtracting the grand mean. In terms of the matrix presented on Table 1 (assuming that the cell entries are now log likelihoods rather than probabilities), the rows are first averaged. Then, the resulting row marginals are averaged to obtain the grand mean. The initial item estimate, $d_i$, is obtained by subtracting the grand mean from the row marginal. The following equation gives the operations necessary to obtain the initial item estimates:

$$d_i = 1/(k - 1) \sum_{j}^{k-1} t_{ij} - 1/k(k - 1) \sum_{i}^{k} \sum_{j}^{k-1} t_{ij}, \tag{4}$$

[4]The cell *probabilities* may be reproduced from the log likelihoods for items and persons. The following equation gives this relationship, where $b_j$, $d_i$, and $P_{ij}$ are defined as before:

$$P_{ij} = \frac{\exp (b_j + d_i)}{1 + \exp (b_j + d_i)}.$$

where $j$ is the subscript for score groups and $i$ is the subscript for items. The first term refers to averaging over the columns to obtain the row marginal, while the second term is the grand mean of the matrix.

The initial values for persons, $b_j$, are obtained in a similar manner, averaging the log likelihoods for each score group over the items and subtracting the grand mean. Referring again to Table 1, the columns are averaged and the grand mean is subtracted from the resulting column marginals. The following equation gives these operations:

$$b_j = (1/k) \sum_{i}^{k} t_{ij} - 1/k(k-1) \sum_{i}^{k} \sum_{j}^{k-1} t_{ij}. \tag{5}$$

The final item and person parameter estimates are determined by the maximum likelihood procedure developed by Wright and Panchapakesan (1969). This procedure simultaneously solves two sets of equations until the estimates converge from one interation to the next. The first condition to be satisfied is maximum predictability of the observed frequencies of passing each item for each score group from the estimated parameters of the model. The second condition is maximum predictability of obtained raw scores from a sum of the predicted probabilities, $P_{ij}$, that the score group will pass each individual item. The final estimated parameters, then, maximize the fit of the model to the data in the item by score group matrix.

*Item Calibration and Unweighted Score Groups*

Whether the model is conceptualized in terms of simple likelihoods or log likelihoods, it is important to notice that each cell in the item by score group matrix has equal weight in determining the initial estimates of the parameters. The observed likelihoods of passing an item are summed over to estimate the initial item easiness parameters, *without respect to the size of the groups obtaining each raw score.* It makes no difference, then, if the estimates come from a high-ability sample, where high scores are obtained more frequently than low scores, or from a low-ability sample, where the reverse is the case. The Rasch model is concerned with reproducing the observed pattern of likelihoods associated with raw score groups. In contrast, traditional item analysis techniques are concerned with the likelihood or probability that a member of a given population can pass an item.

This particular feature of the Rasch model is critical with respect to claims about the invariance of item parameters over different nonrandom samples from the same population. When the specific distribution characteristics of a sample with respect to a latent trait are not permitted to weight the estimates, the item parameters will be sample-free. However, it is important to notice that this is true *only if there is no "interaction effect" between samples and items.* The item parameters will be invariant only if the same likelihoods are associated with items for each score group in each nonrandom sample. The more "culturally-biased" the items, the less likely item parameters will be invariant. In the final analysis, then, sample-invariance of items is a necessary condition for the Rasch model to provide sample-free calibrations.

The shift in emphasis from populations to raw score groups has one important operational implication: huge $N$'s are required. Unlike classical item analysis, each score group is used to give independent estimates of the item parameters. However, even when as many as 500 persons are used for item calibration, extreme scores may

not be obtained frequently enough to provide very stable estimates of the $P_{ij}$'s. Even if scores on a 50-item test formed a perfectly rectangular distribution, for instance, a total $N$ of 500 would produce only about 10 persons per score group. Typically, however, mid-range score groups have very high frequencies and extreme score groups may have few or no observations at all. Although the $P_{ij}$'s from the extremes can be estimated from the model, the need for very large $N$'s during test development should be obvious.

*Anchoring and Interpreting Ability Scores*

The key to the sample-invariant interpretability of ability scores and comparability of ability scores from any item subset is the manner in which scores can be anchored. Whether the model is presented in terms of simple likelihoods or log likelihoods it can be seen that there is an indeterminacy in the solution of the parameter values. With simple likelihoods, for example, the ability parameters can be multiplied by any number, as long as the items are divided by the same number. Similarly, constants may be added or subtracted when the model is expressed as log likelihoods.

The indeterminacy of the ability and item parameters permits the likelihoods to be anchored either to a set of items or a group of persons. To anchor to items, a set of items which have theoretical importance in measurement of the latent trait can be used as the standard set. The anchoring is accomplished by setting the mean simple likelihood for the item set equal to 1.0. The parameters for ability are then adjusted to compensate for the item anchoring. When items are used for the anchoring, ability scores are interpreted relative to the likelihood of solving an item in the standard set. For example, a person performing at the level of the item set (i.e., has a 50/50 likelihood of passing any item) should have an ability score of 1.0.

Similarly, the solution for the Rasch parameters can be anchored to some group of persons. The persons may represent either typical, minimal or optimal ability levels; or the group may be a population of special interest. The average ability likelihood for the group can be set at 1.0 and the item likelihoods adjusted accordingly. Both a person with an ability likelihood of 1.0 and an item with a likelihood of 1.0 would be at the level of the group.

The manner in which anchoring defines the unit of measurement for the Rasch ability parameters may be contrasted to norm-referenced (see Popham & Husek, 1969, for this definition) measurement on traditional ability tests. Objective test score interpretation on traditional tests is made possible through the comparison of an individual's score with some appropriate group. How much ability a person has, by z-scores or percentile ranks, depends on with whom he is compared.

In contrast, anchoring the Rasch ability parameters to a standard set of items leads to domain-referenced interpretations of individual scores. Rather than depending on the trait distribution in various subgroups of persons, item-anchored ability interpretations depend on the theoretical importance or representation of the trait from the standard set of items. As with the more typical usage of domain-referenced interpretations in achievement testing, the ability scores are indices of success on a specified group of items. If the simplest domain-referenced score, percentage correct, is used as an estimate of the ability associated with each raw score, it is easy to see that this score will have the same interpretation regardless of group membership. The interpretability of Rasch ability parameters is, of course, slightly more involved than the direct in-

terpretability of percentage correct scores, differing mainly as to the sophistication with which ability and item information is used. However, the basic rationale is similar.

Anchoring Rasch ability parameters to a standard group of persons seems, on the surface, similar to a norm-referenced approach. However, as with item anchoring, individual score interpretations are sample invariant since the scores are not interpreted as relative standings in a population. As Wright (1967) has indicated, ability likelihoods may be interpreted as direct ratios to the group mean—i.e., person X has twice as much ability as the group average, etc. The need to select a standard group which has intrinsic theoretical importance with respect to the trait should be obvious, if ratio interpretations are to be meaningful.

Unlike either norm-referenced ability tests or domain-referenced achievement tests, however, the anchoring of the Rasch ability parameters means that a person's capability can be estimated by using any subset from the calibrated item pool. The major prerequisite is local independence of items so that each item parameter and error of estimate remain invariant over test content. Once the item parameters are calibrated and anchored on some sample, the Rasch ability equivalents of all raw scores for any set of items may be produced from the standardization data. The comparability of the ability estimates results from fixing the item parameters relative to the likelihoods associated with some standard set of items, rather than the particular subset. Thus, the ability parameters will estimate the likelihoods of passing items in the whole set, rather than the particular subset, which may not represent the difficulty of the whole set.

*Equivalency and Precision*

Wright (1967) suggested that since statistically equivalent forms can be obtained by using any item subset, the use of the Rasch model eliminates the need to painstakingly equate items to create equivalent forms. However, there is quite a difference between equivalent forms in the traditional sense and the kind of statistical equivalency that may be derived from Rasch-calibrated item subsets by the Wright and Panchapakesan (1969) estimation procedure. Lord and Novick's (1968) distinction between parallel measures and tau-equivalent measures helps clarify some of the differences between the traditional meaning of test equivalency and the equivalency of Rasch-calibrated item subsets. Parallel measures, the traditional concept of equivalency, measure the trait equally well for all persons, since the expected values for both true scores and error variances are equal. Parallel tests will then have equal means, variances, and intercorrelations with all other variables. Tau-equivalent measures, however, have equal expected values for true scores, but not necessarily equal error variances. Under this more limited definition, equivalent forms need not have equal variances or reliabilities. The equivalency of Rasch item subsets falls under the more limited definition of equivalent measures.

Another difference of Rasch-calibrated item subsets from classical parallel measures is the relationship between test equivalency and the precision of true score estimates. In the classical model, the same index, correlation between alternative test forms, is often used as evidence for both test equivalency and measurement precision. Furthermore, internal consistency reliability, another commonly used index of precision, is also related to test equivalency in the following ways: (1) the internal consistency of equivalent forms should be equal and (2) the degree of precision estimated by alternative form correlations of equivalent tests should not be substantially lower

than the respective internal consistencies. As will be shown in this section, the test equivalency of Rasch-calibrated item subsets does not show the same relationship to precision as for classical parallel measures. The main difference derives from the different indices of error and precision which are permitted by the local independence of subjects and items in the Rasch model.

Wright and Panchapakasen (1969) apply the binomial model to estimate error, separately for each item and score group (ability) parameter. The standard error of estimate for items is approximated by the following equation:

$$v(d_i) \doteq (1/[k - 1]^2) \sum_{j=1}^{k-1} 1/(r_j P_{ij}[1 - P_{ij}]), \tag{6}$$

where probability of correct response $P_{ij}$ = as estimated by parameters for cell $ij$, $r_j$ = number of persons obtaining raw score $j$, and $\doteq$ indicates approximately equal. It can be seen that the standard error of the item becomes small as the term $[r_j P_{ij} (1 - P_{ij})]$ increases. This term is maximized when the probability of passing the item is as close as possible to the probability of failing for the maximum number of persons in the standardization sample. For a sample with a normal distribution of ability scores, the standard error of the item will take on its smallest value when the probability of passing the item is .50 for the sample as a whole. Although the item easiness parameter estimate is sample-invariant, item error in the Rasch model is sample-specific.

Unlike classical test models, where measurement error is assumed to be equal for all ability levels as a latent trait model, separate errors of measurement can be provided for each Rasch-scaled ability level. The standard error of an ability estimate is approximated by the following formula:

$$v(b_j) \doteq (1/k^2) \sum_{i=1}^{k} 1/r_j P_{ij}(1 - P_{ij}). \tag{7}$$

That is, the inverse of the predicted cell frequencies are summed over items and then multiplied by $1/k^2$ to give the standard error of the ability estimate. As with items, the standard error is minimized for a score group when for as many cells as possible the probability of passing equals the probability of failing. Also, score groups with larger frequencies, $r_j$, will have smaller standard errors than those with fewer persons. Over all score groups, the standard error will decrease as the number of items increases since $(1/k^2)$ will be minimized.

It can be seen, then, that the precision of estimating ability for any particular score group depends on which items are used. The most precise ability estimate for a score group occurs when as many items as possible are at the 50% difficulty level for the group. Following this line of reasoning, the best item subsets to use for different populations will vary when these populations vary with respect to the latent trait, if ability is to be estimated with maximum precision for the population as a whole.

Regardless of the average size of measurement error for a group, Wright (1968) claims that the observed difference in estimation between any calibrated item subsets will be totally accounted for by the associated measurement errors. That is, the differences between ability scores on the two test forms will be distributed as would be expected from the confidence intervals associated with the scores obtained on each

test. To make a test of statistical equivalency, a "standardized difference score" must be computed for each person. This is given by the following formula:

$$D_{12} = \frac{x_{1p} - x_{2p}}{(SE^2_{x1p} + SE^2_{x2p})^{1/2}}, \tag{8}$$

where $D_{12}$ = standardized difference, $x_{1p}$, $x_{2p}$ = ability score obtained by person $p$ on test 1 and test 2 respectively, and $SE^2_{x1p}$, $SE^2_{x2p}$ = measurement errors associated with $x_{1p}$, $x_{2p}$. The observed difference between the ability estimates given by the two tests is divided by the standard error of the score differences. The standardized difference score computed for each person can be interpreted as a z score, which compares his observed ability difference to the theoretical distribution of the differences expected from the measurement error associated with each score. If the error between the two tests is accounted for by measurement error, then when the standardized differences are summed over persons in the population, these scores should be normally distributed with a mean of 0 and standard deviation of 1.0.

Statistical equivalency of any item subsets, as obtained from applications of the Rasch logistic model, merely means that the observed ability differences between subset scores are distributed as would be expected from measurement error alone. However, there is no guarantee that "statistically equivalent forms" are also maximally precise forms. Precision of measurement, as shown above, is still population-specific and will not be maximized unless items are carefully selected. Similarly, to minimize error in predicting ability scores from one test to the other, it is not possible to use just any subsets of items from the calibrated pool since the measurement error on each form must be minimized.

## EQUIVALENCY OF CALIBRATED ITEM SUBSETS

Tinsley (1971) compared the equivalency of item subsets on four tests and concluded that the ability estimates from the Rasch model were not invariant over item subsets. However, Tinsley did not use standardized differences in his comparisons and confounded precision with statistical equivalency. Data from one of Tinsley's tests were reanalyzed for two reasons: (1) to compare the observed ability differences between item subsets to the theoretical distributions of measurement error for scores, and (2) to examine the relative degree of precision of measurement between subsets.

### *Procedure*

Test protocols for a 60-item verbal analogies test were calibrated by the Wright and Panchapakesan (1969) procedure for estimating Rasch model parameters. All items were multiple-choice, with five alternatives. The items on this test had been selected from a group of 96 items which were administered to college students. The items had been selected according to mixed criteria, with fit of the item data to the Rasch model as one of these criteria.

Data from 949 subjects were available on the final 60-item analogies test. Approximately two-thirds of the sample were college students, while the remaining one-third of the sample consisted of suburban high school students. The 60-item test had a mean of 34.86, and a variance of 89.32, and the Hoyt reliability coefficient was .877, showing an adequate degree of internal consistency for an analogy test. However, 30%

of the items did not fit the Rasch model at the .01 level, while 40% of the items did not fit when the more stringent criterion of .05 was used.[5] Thus, the robustness of the model with respect to equivalent forms was tested with the anlaogy test, since several items do not fit the model and the multiple-choice format yields possible guessing bias.

Three different divisions of the pool of 60 calibrated items resulted in the following subset comparisons: (1) odd versus even items, (2) easy versus hard items, and (3) randomly selected subsets with no item overlap. Each subset contained 30 items. The corresponding log ability estimates for obtained raw scores on each subset were obtained by a maximum likelihood procedure for each subset using the fixed item parameters, estimated from the full 60-item calibration on 949 subjects.

*Results*

Table 2 presents the means and variances for both log likelihood and raw scores on the six item subsets. The results from the comparisons between item subsets indicated that the raw score means and variances differed widely. For all three subset comparisons, the means were significantly different ($p \leq .05$). The odd-even and easy-hard subsets were significantly different in variability, while the random subsets did not differ. The $t$ values reported are for correlated variances (Guilford, 1956).

Scaling the test in log likelihoods produced fewer significant differences between subsets. The only significant differences were between the easy and hard item subsets, which had both significantly different means and variances. Although the mean difference, in absolute terms is probably too small to be theoretically important, the difference in variability is sizable.

Table 2 also presents on the precision of the Rasch-scaled ability estimates for the sample as a whole. The individual score errors were weighted by the relative frequency of the scores to obtain the mean standard error of measurement. Table 2 shows that the average expected measurement error is approximately equal between the random

Table 2

Means, Variances and Measurement Errors of Item Subsets for Log Likelihood and Raw Scores

| Subset | % Pass | Hoyt Index | Raw Scores | | | | Log Likelihood | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | $\bar{x}$ | $t_{\bar{x}_1-\bar{x}_2}$ | $s^2$ | $t_{s^2}$ | $\bar{x}$ | $t_{\bar{x}_1-\bar{x}_2}$ | $s^2$ | $t_{s^2}$ | $SE_{msmt.}$ |
| Odd | .65 | .77 | 19.38 | 17.78 | 23.67 | 2.99 | .432 | .86 | .873 | .51 | .453 |
| Even | .52 | .78 | 15.47 | | 26.85 | | .415 | | .892 | | .433 |
| Easy | .74 | .83 | 22.31 | 42.52 | 29.43 | 6.23 | .469 | 2.15 | 1.178 | 13.76 | .503 |
| Hard | .42 | .72 | 12.53 | | 22.18 | | .415 | | .663 | | .419 |
| Random Set I | .59 | .77 | 17.79 | 2.86 | 24.25 | .69 | .427 | .43 | .882 | .67 | .447 |
| Random Set II | .57 | .77 | .7.06 | | 25.83 | | .436 | | .857 | | .433 |

[5]The .01 and .05 significance levels refer to the probability of the observed item data, given that the item fits the model. The smaller the probability, the less the observed item data conforms to the distribution specified by the Rasch model.

Table 3

Precision of Measurement and Standardized Differences Between Item Subsets

| Subset Comparison | Log Likelihood | | Standardized Difference | | |
|---|---|---|---|---|---|
| | $s^2_{\bar{x}_1-\bar{x}_2}$ | r | $\bar{x}_1-\bar{x}_2$ | $s^2_{\bar{x}_1-\bar{x}_2}$ | $s_{\bar{x}_1-\bar{x}_2}$ |
| Odd vs. Even | .425 | .76 | .007 | 1.028 | 1.014 |
| Easy vs. Hard | .590 | .76 | -.057 | 1.313 | 1.146 |
| Random Sets | .410 | .76 | -.020 | .995 | .998 |

subsets and the odd-even subsets, but that the precision of measurement is greater for the hard test than the easy test.

A somewhat different pattern of results is obtained from a classical model index of reliability, internal consistency. The Hoyt index of reliability was approximately equal between the random subsets and the odd-even subsets, but differed between the hard and easy subsets. However, in direct contrast to the estimated precision for Rasch-scaled ability scores, these results show that the easy test has the least measurement error.

Table 3 presents two types of data on the equivalency of the item subsets. Standardized difference scores, used to compare subset equivalency for Rasch-scaled estimates, were computed for each of the three comparisons. The mean standardized difference scores did not differ significantly from zero for all three comparisons, as would be expected for equivalent subsets. The variances were very close to the expected values of 1.0 for equivalent forms for both the random subsets and the odd-even comparison. The easy versus hard test comparison, however, yielded standardized difference scores with a variance significantly larger than 1.0 ($F = 1.3, p < .01$). The other type of equivalency data came from the classical model, correlations between subsets. Table 3 shows that the parallel form correlations were uniform for the three comparisons, but were only moderately high. Similarly, Table 3 shows that the variance of the differences between subsets (uncorrected for measurement error) is substantial.

*Discussion*

The results from two of the subset comparisons, odd-even and random sets, demonstrate the statistical equivalency of Rasch-calibrated item subsets which have not been matched for difficulty. The standardized differences between these subsets were distributed as would be expected from measurement error alone. These results

are notable because the classical model requirements of test equivalency were not fully met for either comparison. Test means were unequal for both the odd-even and random sets comparisons, while the variances were unequal for the odd-even comparison.

However, the results from a third comparison, hard versus easy items, indicated differences which could not be accounted for by the measurement error estimated for the Rasch ability scores. Although the reason for this difference is not entirely clear, it is possible that poor fit of items to the model may have been a factor. To determine the plausibility of this interpretation, the percentage of items not fitting the model on the separate subsets was computed. It was found that 23% of the items on the easy subset and 57% of the items on the hard subset did not fit the model. Thus, both ability and measurement error were inaccurately estimated on the hard subset, since the responses to many of the difficult items did not conform to the item characteristics curve specified by the logistic model.

The poor fit of the difficult items to the model also explains the inconsistent results on precision from internal consistency versus the Rasch ability error. Although the hard test theoretically provides the more precise measurement (that is, the probability of passing the items was closer to .50 for the most persons), many of the items apparently were not highly correlated with total score, causing both the poor fit to the Rasch model and the lower internal consistency.

In general, these results indicate that only under the most extreme conditions does the Rasch model fail to produce statistically equivalent forms for any item subsets. However, none of the item subsets resulted in the equivalency characteristic of tests developed by classical techniques. Furthermore, some increase in precision could have been gained by more efficient item selection, as evidenced by the varying average measurement error between forms. The Rasch ability error estimate and an index of internal consistency yielded different results, showing the need to consider both the slope of the item characteristic curve and the appropriateness of the difficulty level for subjects in this particular set of items.

## CONCLUSION

A thorough analysis of applying the Rasch logistic model to trait test data revealed that objective measurement results from a complex interaction of the properties of the model, the nature of the estimation technique and the characteristics of the test data. Applying the Rasch model to typical trait data does not necessarily yield objective measurement, since some of the claimed advantages of applying the model depend directly on the characteristics of the item pool, rather than the model. For an item pool fully to possess the properties of objective measurement, a set of rigorous conditions must be met.

The most direct influence of item characteristics is on the sample-invariance of item calibrations. The Rasch item parameter estimates will be invariant only under a special condition. Not only must individuals with the same raw score have the same probabilities of passing each item, regardless of the sample to which they belong, but the item characteristics curves must have equal slopes. Thus, item parameters will not be sample-invariant when there is cultural bias which differentially affects the item probabilities and alters the slopes. Since it is well known that many popular ability

tests have items which differ in cultural loadings, the special condition required for item parameter invariance may be difficult to obtain.

The difficulty in obtaining suitable items for the Rasch model, however, should not diminish its *theoretical* superiority over classical approaches to test development. The traditional index of item difficulty, percent passing, is always a population-specific statistic, determined by the distribution of the trait in a particular population. To adequately estimate item difficulty, samples must be randomly selected. In contrast, the Rasch item parameter is independent of the trait distributions in a particular population and the estimations need not be from random samples, if samples and items do not interact. Although currently available test data may not meet this requirement, future refinements in substantive measurement theory and techniques may extend the practical applicability of the Rasch model.

Another claimed advantage of using the Rasch model, sample-invariant interpretations of ability scores, is probably obtainable for many item pools. Unlike traditionally constructed tests, the interpretation of the anchored Rasch ability parameters does not depend on the distribution of the trait in specific populations. Anchoring the solution for the Rasch parameters to a set of items or a group of persons, eliminates the need for numerous norm tables to objectively interpret individual scores. However, the anchored score interpretations will be only superficially objective unless either explicit item-trait theory or a group of substantial interest to trait theory are available. The difficulty in obtaining a meaningful interpretation for the anchored parameters is similar to the interpretation problems in domain-referenced testing; the status of substantive knowledge in measurement is, at best, inadequate. Although use of the Rasch model offers a new basis for score interpretability, current trait theory is not sufficiently advanced to permit wide applicability.

A major focus of this paper has been on the construction of equivalent forms from a calibrated item pool. Use of the Rasch model was found to have many more parallels to traditional criteria for the development of equivalent forms than would have been anticipated from previous explanations (Wright & Panchapakesan, 1969). To compare Rasch-calibrated item subsets to traditional equivalent forms, it was found necessary to consider statistical equivalency independently from precision. Statistical equivalency, in the narrow sense, means that the ability difference scores obtained by a sample are distributed as would be expected from the measurement error associated with each score. Precision, however, means that the measurement differences between tests are as small as possible. In theory, any two subsets from an item pool calibrated with the Rasch model will yield statistical equivalency but not necessarily maximum precision. In comparison, traditional equivalent forms maximize both precision and statistical equivalency.

The empirical results generally substantiated the theoretical interpretation of the nature of equivalent forms from the Rasch model. Only under extreme conditions did the measurement errors fail to account for the observed differences between subsets. However, none of the subsets were equivalent in the traditional sense. Alternate form correlations were only moderate, and there was some evidence that precision might have been increased by using more efficient techniques in selecting items. Although the Rasch item parameter may be invariant over populations, precision is specific to the trait distribution in a given population. If the goal of item selection is to develop fixed-

content tests, then the classical techniques of having item difficulties close to .50 and matching extreme item difficulties will yield the most precise equivalent forms.

A primary strength of the statistical equivalency of Rasch-calibrated item subsets, however, is the possibility of individualized selection of items rather than the construction of fixed-content tests. Although the use of the Rasch model cannot improve precision in fixed-content tests, the special properties of a latent trait model permit the desired degree of precision for any person to be obtained from the fewest possible items. If the test data meets the latent trait model assumption of independence of items and persons, measurement error can be estimated separately for each score level. Furthermore, the anchoring of the items to a standard set permits the estimation of ability and measurement error for any subset of items, without standardizing the subset on a new sample. If items are administered by a computer, ability and measurement error can be estimated after the person responds to each item. The next item selected, then, can be as close to the ability estimate as possible and will give the largest increase in precision.

It is important to realize that the assumption of local independence of items is critical for individualized testing. The administration of a given item must not influence the subject's responses to following items. When items are not locally independent, the subject's performance, in part, will depend on which items he has already completed. In this situation, only fixed content tests will provide comparable estimates of the subject's ability. The degree to which test items interact will have to receive systematic study if individualized testing is to be successful.

Specifying measurement error for each score level, rather than the test as a whole, has an additional advantage; ability change at different score levels may be compared on a comparable statistical basis. A typical trait test is not equally precise for all populations. Extreme scoring populations can be expected to change more than mid-range scoring populations because of the greater measurement errors for populations at the tails of the total distribution. The standardized difference score (a z-ratio computed by dividing the differences in Rasch ability estimates by the measurement error associated with each score) permits comparison of change at different ability levels, since ability change is adjusted for the individual measurement errors.

In conclusion, the lack of impact of the Rasch model in test development is due more to the current status of trait measurement than to the properties of the model. Many of the advantages of the Rasch model necessitate a different kind of data for trait measurement than is now characteristic of the field. Explicit trait-item theory, locally independent items and routine administration of tests by computer, would be part of the necessary technological sophistication.

## REFERENCES

ANDERSON, J., KEARNEY, G. E., & EVERETT, A. V. An evaluation of Rasch's structural model for test items. *The British Journal of Mathematical and Statistical Psychology*, 1968, **21**, 231–238.

BROOKS, R. D. *An empirical investigation of the Rasch ratio-scale model for item difficulty indexes.* Doctoral dissertation, University of Iowa, Ann Arbor, Michigan: University microfilms, 1965, No. 65–434.

GUILFORD, J. P. *Fundamental statistics in psychology and education.* New York: McGraw-Hill, 1956.

LORD, F. M., & NOVICK, M. R. *Statistical theories of mental test scores.* Reading, Massachusetts: Addison-Wesley Publishing Company, 1968.

PANCHAPAKESAN, N. The simple logistic model and mental measurement. Unpublished Doctoral dissertation, University of Chicago, 1969.

POPHAM, W. J., & HUSEK, T. R. Implications of criterion-referenced measurement. *Journal of Educational Measurement,* 1969, **6,** 1–9.

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. In *Proceedings of the fourth Berkeley Symposium on mathematical statistics.* Berkeley: University of California Press, 1961, IV, 321–334.

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

RASCH, G. An individualistic approach to item analysis. In P. F. Lazarsfeld and N. W. Henry (Eds.), *Readings in mathematical social science.* Chicago: Science Research Associates, 1966, Pp. 89–108. (b)

TINSLEY, H. E. A. An investigation of the Rasch simple logistic model for tests of intelligence or attainment. Unpublished doctoral dissertation, University of Minnesota, 1971.

WRIGHT, B. Sample-free test calibration and person measurement. *Proceedings of the 1967 Invitational Conference on Testing Problems.* Princeton, N.J.: Educational Testing Service, 1967. Pp. 85–101.

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

## AUTHORS

WHITELY, SUSAN E. *Address:* Psychology Department, University of Kansas. *Title:* Assistant Professor of Psychology. *Degrees:* B.A., Ph.D. University of Minnesota. *Specialization:* Measurement.

DAWIS, RENÉ V. *Address:* Psychology Department, University of Minnesota. *Title:* Professor of Psychology. *Degrees:* B.A. University of the Philippines, M.A., Ph.D. University of Minnesota. *Specialization:* Counseling psychology and measurement.