DOCUMENT RESUME

ED 075 484 TM 002 562

AUTHOR Whitely, Susan E.; Dawis, Rene V.
TITLE The Nature of Objectivity with the Rasch Model.
INSTITUTION Minnesota Univ., Minneapolis. Center for the Study of Organizational Performance and Human Effectiveness.
SPONS AGENCY Office of Naval Research, Washington, D.C. Personnel and Training Research Programs Office.
REPORT NO TR-3008
PUB DATE 2 Jan 73
NOTE 29p.

EDRS PRICE MF-$0.65 HC-$3.29
DESCRIPTORS *Equivalency Tests; *Item Analysis; *Mathematical Models; *Measurement Instruments; Psychometrics; Statistical Studies; Technical Reports; *Test Construction
IDENTIFIERS *Rasch Model

ABSTRACT

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 along 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 relation of underlying assumptions and computations in the Rasch model to trait theory and trait measurement has 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. (Author)

THE CENTER FOR THE STUDY OF ORGANIZATIONAL PERFORMANCE AND HUMAN EFFECTIVENESS

University of Minnesota
Minneapolis, Minnesota

Office of Naval Research Contract
ONR N00014-68-A-0141-0003

Prepared for

PERSONNEL AND TRAINING RESEARCH PROGRAMS
PSYCHOLOGICAL SCIENCES DIVISION
OFFICE OF NAVAL RESEARCH

Contract No. 00014-68-A-0141-0003
Contract Authority Identification Number, NR No. 151-323

THE NATURE OF OBJECTIVITY WITH THE RASCH MODEL

Susan E. Whitely and Rene' V. Dawis

Technical Report No. 3008

Security Classification

DOCUMENT CONTROL DATA - R & D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY (Corporate author) The Center for the Study of Organizational Performance and Human Effectiveness University of Minnesota, Minneapolis, Minnesota 55455	2a. REPORT SECURITY CLASSIFICATION UNCLASSIFIED 2b. GROUP

3. REPORT TITLE

The Nature of Objectivity with the Rasch Model

4. DESCRIPTIVE NOTES (Type of report and inclusive dates)

Technical Report No. 3008

5. AUTHOR(S) (First name, middle initial, last name)

Susan E. Whitely and Rene' V. Dawis

6. REPORT DATE	7a. TOTAL NO. OF PAGES	7b. NO. OF REFS
2 January 1973	25	11
8a. CONTRACT OR GRANT NO. N00014-68-A-0141-0003 b. PROJECT NO. NR 151-323 c. d.	9a. ORIGINATOR'S REPORT NUMBER(S) 3008 9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report)	

10. DISTRIBUTION STATEMENT

Approved for public release; distribution unlimited

11. SUPPLEMENTARY NOTES	12. SPONSORING MILITARY ACTIVITY
	Personnel and Training Research Programs Office of Naval Research Arlington, Virginia 22217

13. ABSTRACT

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 along 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 relation of underlying assumptions and computations in the Rasch model to trait theory and trait measurement has 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.

DD FORM 1 NOV 65 1473 (PAGE 1)

S/N 0101-807-6811

Security Classification

14 KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Rasch model Test equivalence Objective measurement Test development						

DD FORM 1 NOV 65 1473 (BACK)
S/N 0101-807-6821

Security Classification

A-31409

The Nature of Objectivity
With the Rasch Model

Susan E. Whitely and Rene' V. Dawis

A new kind of item analysis, originally formulated by Rasch (1960, 1966a, 1966b), is now available for use in developing measures of unidimensional traits. Wright (1968), 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. According to Wright (1968), tests calibrated by the Rasch model will have the following characteristics: 1) the calibration of the measuring instrument is independent of the sample and 2) the measurement of a person on the latent trait is independent of the particular instrument used. A psychological test having these objective 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 the Rasch model has the following properties: 1) the estimates of the item difficulty parameter will not vary significantly over different populations of people, 2) the estimates of a person's ability, given a certain raw score, will be invariant over different populations and 3) estimates of a person's ability from any calibrated subset of items will be statistically equivalent. If these properties are truly characteristic of the Rasch technique, 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 subset of the calibrated item pool could be used as alternative instruments. Similarly, independence of measurement from a particular population norm 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 technique has had little apparent impact. No major attempt at test development has yet been reported. The reasons for this are not clear, particularly since initial research has been encouraging. Both item and ability parameters have been found to be population-invariant (Anderson, Kearney and Everett, 1968; Brooks, 1965; Tinsley, 1971). Furthermore, the model appears to be robust with respect to several of the underlying assumptions (Panchapakesan, 1968). However, little evidence on the equivalency of item subsets has been presented, nor is it clear from Wright and Panchapakesan's (1969) paper how the model accomplishes either item-invariance or population-invariance of the estimated parameters.

The major purpose of the present paper is to determine how the Rasch model's underlying assumptions, computational procedures and trait theory interact to produce item- and population-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) in the cells of an item-by-score-group matrix, in which 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 or probabilities by use of only two parameters, item easiness and person ability, in the following manner:

$$(1) \quad P_{ij} = \frac{A_j \times E_i}{1 + A_j \times E_i}$$

where A_j = ability parameter for score group j

and E_i = easiness parameter for item i

<u>Assumptions</u>. The most basic assumption made by the Rasch model is unidimensionality of the item pool. If subjects are grouped according to total score, within each group there should be no remaining significant correlations between items. This means that 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.

A second assumption, 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. It is assumed that both of these ordering conditions will hold true for any population, regardless of the mean value of the latent trait.

Two additional assumptions must be made in order to apply the simple logistic model proposed by Rasch. 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. Also, there must be minimal guessing so that the probability of passing an item by chance is minimized.

As summarized by Wright and Panchapakesan (1969), it is assumed that the only way in which items differ is in 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 assumptions (Anderson, Kearney and Everett, 1968; Panchapakesan, 1969; Wright and Panchapakesan, 1969). However, as will be pointed out in this paper, the population-invariance feature of the Rasch model with respect to item calibration is actually an assumption, and the amount of departure from this feature depends directly on the degree to which there is an "interaction effect" between populations and items.

Estimating the parameters. An understanding of how the item and person parameters are determined necessitates converting the cell probabilities into likelihood ratios.[1] Likelihood ratios are simply betting odds, the ratio of the probability of passing to the probability of failing. In terms of likelihoods, the cells are to be reproduced by the simple product of item easiness and person ability values as follows:

$$(2) \quad \frac{P_{ij}}{1 - P_{ij}} = A_j \times E_i$$

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 an item in the set, whereas the item easiness value indicates the likelihood the item will be passed. How these likelihoods are derived constitute the major concern in this section.

The initial values for ability and easiness are directly derived from the values in the corresponding row or column. Item easiness is estimated by the k-1 root of the product of the score group likelihoods, as follows:[2]

$$(3) \quad E_i \approx \sqrt[k-1]{\left(\frac{P_{i1}}{1 - P_{i1}}\right)\left(\frac{P_{i2}}{1 - P_{i2}}\right) \cdots \left(\frac{P_{i,k-1}}{1 - P_{i,k-1}}\right)}$$

where k-1 = the number of score groups

Thus, the item parameters are initially estimated by the geometric mean of the likelihoods across score groups. The comparable initial values for person ability can be similarly obtained by taking the geometric mean across items, as follows:

$$(4) \quad A_j \approx \sqrt[k]{\left(\frac{P_{1j}}{1 - P_{1j}}\right)\left(\frac{P_{2j}}{1 - P_{2j}}\right) \cdots \left(\frac{P_{kj}}{1 - P_{kj}}\right)}$$

Thus, it can be seen that the initial ability estimate for a score group is the "average" likelihood of passing an item in the set.

Both the initial values and the final values are usually reported as log likelihoods rather than simple likelihoods. The log likelihood for an item easiness estimate, d_i, is simply the arithmetic mean of the log likelihoods over score groups, as follows:

$$(5) \quad d_i = \frac{\Sigma t_{ij}}{k-1}$$

where t_{ij} = cell log likelihoods

and, of course, the antilog of this value is E_i. Similarly, the arithmetic mean of the log likelihoods over items estimate the log ability estimates, b_j, as follows:

$$(6) \quad b_j = \frac{\Sigma t_{ij}}{k}$$

The log likelihood scale for item easiness and person ability is related to the probability a score group will pass an item, as given by the following equation:

$$(7) \quad P_{ij} = \frac{\exp\ (b_j + d_i)}{1 + \exp\ (b_j + d_i)}$$

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 is the simple sum of log A_j and log E_i as follows:

$$(8) \quad t_{ij} = b_j + d_i$$

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.

A computational step in the model which is important in the final interpretation is the anchoring of the parameters. Since item and person parameters are conjointly estimated from the same function, a unique solution is not specified. To provide an anchor for the item easiness estimates, the mean of the item log likelihoods is set equal to zero by subtracting the grand mean of the matrix, $t_{..}$, as follows:

$$(9) \quad \log E_i = d_i = t_{i.} - t_{..}$$

Similarly, the person ability estimates must also be adjusted to correspond to the anchoring of the item easiness estimates by setting the mean log likelihood for ability equal to zero as follows:

$$(10) \quad \log A_j = b_j = t_{.j} - t_{..}$$

Thus, as with items, the grand mean is subtracted from the parameter estimated for each score group.

In terms of simple likelihoods, both the mean item easiness likelihood and mean person ability likelihood is set at 1.0. The importance of this anchoring will become clear in the discussion on precision of item subsets.

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 iteration 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. This is given by the following equation:

$$(11) \quad a_{ti} = \sum_{j=1}^{k-1} (r_j \exp [b_j + d_i])/(1 + \exp [b_j + d_i])$$

where r_j = number of persons in score group j

and a_{ti} = number of persons passing item i

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. This condition is given by the following equation:

$$(12) \quad j = \sum_{i=1}^{k} (\exp [b_j + d_i])/(1 + \exp [b_j + d_i])$$

where j = raw score for score group

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 population, where high scores are obtained more frethan low scores, or from a low-ability population, where the reverse ...se. 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 populations. When the specific characteristics of a population with respect to a latent trait are not permitted to weight the estimates, the item parameters will be population-free. However, it is important to notice that this is true only if there is no "interaction effect" between populations and items. The item parameters will be invariant only if the same likelihoods are associated with items for each score group in different populations. The more "culturally-biased" the items are, the less likely item parameters are to be invariant over populations. In the final analysis, then, population-invariance of items is an assumption of the model.

The shift in emphasis from populations to 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 no more than 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 population-invariant interpretability of ability scores and to item-invariant equivalency of forms is the manner in which scores are anchored. The subtraction of the grand mean during the computation of the initial item easiness estimates results in the standardization of the item set to a mean likelihood value of 1.0. Ability estimates are correspondingly adjusted such that a person performing at the mean level of the item set would have an ability of 1.0. When the parameters are anchored in this way, ability scores can be interpreted as the odds the person will pass an item in the calibrated set.

The claimed advantages of using Rasch ability parameters rather than the more traditional z-scores or percentiles actually derives from the use of this "domain-referenced" rather than the usual "norm-referenced" interpretation of test scores (cf. Popham and Husek, 1969, for this distinction). 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 1) will have the same interpretation regardless of what population the individual belongs to, and 2) estimates ability on a ratio scale since the zero point can be interpreted as not passing any items. The population-invariant interpretability of Rasch ability parameters is only slightly more involved than the direct interpretability of percentage correct scores, differing mainly as to the amount of information used to derive the ability estimates.

Unlike percentage correct scores, however, the anchoring of the Rasch ability parameters on the item set means that a person's ability can be estimated by using any subset from the calibrated item pool. The major prerequisite is that the item parameters' errors of estimate are known by simultaneously calibrating all the items on some population. These values are then fixed, and the ability associated with each of the possible k-1 scores for any set of k items can be estimated by maximizing the predictability of these raw scores from a sum of the estimated probabilities of passing items for each score group, equation (6). The equivalency of item subsets results from the item parameters being fixed relative to the likelihoods associated with the whole 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.

To compare these instantaneously equivalent forms to those obtained under the more painstaking traditional techniques, three important differences must be noted. The first is that the goals of estimation are limited in the Rasch model. What is being estimated is not some abstract "true" score; rather, ability is defined as the likelihood of solving items in some pre-defined set. The second difference from traditional techniques is that items which fit the Rasch model differ only on difficulty level. Classical item techniques for constructing equivalent forms allow items to differ on other characteristics, such as slope or discrimination. The third, and perhaps most important difference, is the precision with which ability is estimated. This will be considered more fully in the following section.

Precision of measurement. Wright (1968) 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 on tests to create equivalent forms. However, there is quite a difference between statistically equivalent forms in the traditional sense and the narrow kind of statistical equivalency that may be obtained from Rasch-calibrated item subsets.

A claim of statistical equivalency between Rasch-calibrated item subsets merely means that the difference in ability estimation between the forms is no greater than would be expected from measurement error. In contrast, the traditional kind of statistical equivalency results in alternate forms being what might be called "maximally equivalent". The correlations between the test forms are as high as possible so that the precision of ability estimates from one form to the other is maximized. How errors of measurement are estimated in the Rasch model, and what this implies for statistically and maximally equivalent forms, are the major concerns in this section.

For each item and score group (ability) parameter there is an error associated with the estimate. The standard error of estimate for items is approximated by the following equation:

$$(13) \qquad \nu(d_i) \approx (1/[k-1]^2) \sum_{j=1}^{k-1} 1/(r_j\, P_{ij}[1-P_{ij}])$$

where probability of correct response P_{ij} = as estimated by parameters for cell ij

It can be seen that the standard error of the item becomes small as $r_j\, P_{ij}\, (1-P_{ij})$ increases. Given equal frequencies in the score groups, this term is maximized when the probability of passing the item is as close as possible to the probability of failing the item for each score group. Obviously, the difficulty level of the item will increase as a score group's total raw score decreases. So, the standard error of the item will take on its smallest value when the probability of passing the item is .50 for the score group with the largest frequency, r_j. The correspondence to the classical test approach of selecting items with a difficulty level of .50 for the population (to maximize reliability) should be obvious, if the mode and mean of the distribution are equal. So, item error in the Rasch model turns out to be population specific.

Unlike classical test models, where measurement error is assumed to be equal for all ability levels, the Rasch model provides separate errors of measurement for each ability level. The standard error of an ability estimate is approximated by the following formula:

$$(14) \quad \nu(b_j) = (1/k^2) \sum_{i=1}^{k} 1/r_j \; P_{ij}(1-P_{ij})$$

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 be smallest when 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.

Regardless of the average size of measurement error for a group, Wright (1968) claims that the observed difference in estimation between any 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:

(15) $$D_{12} = \frac{x_{1p} - x_{2p}}{\sqrt{SE^2_{x1p} + SE^2_{x2p}}}$$

where D_{12} = standardized difference

x_{1p}, x_{2p} = ability score obtained by person p on test 1 and test 2 respectively

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 of his observed difference between item subset scores on a distribution of the differences that would be expected from the measurement error associated with each score. If the error between the two forms is random, 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, then, merely means that the observed differences between subset scores are distributed as would be expected from measurement error alone. However, even if this claim can be substantiated for item pools calibrated by the Rasch technique, there is no guarantee that statistically equivalent forms are also maximally equivalent forms. The problem of precision, as shown above, is still a population-specific problem. To have "maximally equivalent forms" the measurement error between forms must be minimized and it is not possible to use just any subset of items from the calibrated pool. Items must be as carefully selected as in classical techniques of test development. In fact, the same criteria must be met. Average item difficulty should be at .50, and the test means and variances should be equal if the average standard error of estimate, weighted by frequency, is to be minimized over score groups for each test.

Equivalency of Calibrated Item Subsets

Tinsley (1971) compared the equivalency of item subsets on four tests and concluded that the Rasch ability estimates were not invariant over item subsets. However, Tinsley did not use standardized differences in his comparisons and confounded maximal equivalency with statistical equivalency. Data from one of Tinsley's tests were re-analyzed to determine how well the observed differences between item subsets are accounted for by the errors of measurement for each score and the relative degree of precision of measurement between subsets.

Procedure. Test protocols for a 60-item verbal analogies test were calibrated by the Rasch technique. 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 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 on the combined sample. Hoyt reliability was found to equal .877, showing a good degree of internal consistency in the item pool. However, 30% of the items did not fit the model at the .01 level, while 40% of the items did not fit when the more stringent criterion of .05 was used. Thus, the claims with respect to equivalent forms were to be given a stringent test, since several items do not fit the model.

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, then, contained 30 items. The corresponding ability estimates for obtained raw scores on each subset were obtained by a maximum likelihood procedure using fixed item parameters for each subset. The item parameters were 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. 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 that were found 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 the mean standard errors associated with log likelihood ability estimates, weighted by the frequencies with which this population obtained the various total scores. This error associated with the score groups is approximately equal for the random sets and odd-even comparisons, but does differ between the easy and hard subsets. Apparently the estimated item likelihoods more closely approximate the ability in this population on the hard test than on the easy test. That is, on the hard test, the probability of passing an item is closer to .50 for more score groups than on the easy test. The hard test, then, should provide the more precise measurement for this population.

Table 3 presents more information relevant to the precision of measurement. It can be seen that although the subset mean differences are very small, there is a large variance between the tests. The correlations between the subsets show that the largest percentage of variance shared is only 58% (r=.76). Thus, none of the item subsets are maximally equivalent.

Table 3 also presents the standardized difference for the three comparisons. In no case are the means significantly greater than zero. The variances are very close to 1.0 for both the random sets and odd-even comparisons, but are somewhat larger for the easy versus hard test comparison.

This variance is significantly different from 1.0 (F=1.3, p<.01) and is large enough to have some theoretical importance.

Discussion. The results from two of the subset comparisons, odd-even and random sets, support the claim of statistical equivalency between item subsets calibrated by the Rasch technique. The standardized differences between these subsets were distributed as would be expected from measurement error alone. The results from the standardized differences between the easy and hard subsets, however, indicate that these differences cannot be fully explained by estimated measurement error. Although the reason for this difference is not entirely clear, it is quite likely that the large number of items not fitting the model was the major influence. 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. It is likely, then, that both ability and measurement error were underestimated on the hard subset, since many of the difficult items did not adequately measure the person's ability.

Apparently 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 maximal equivalency characteristic of tests developed by classical techniques, since the correlations between subsets were only moderate. Some increase in precision could have been gained by more efficient item selection, as evidenced by the varying average measurement error between forms. The variance of the Rasch ability estimates was significantly different between the easy and the hard item subsets. The more extreme estimates were obtained from the easy subset, as would be expected, when the population is of relatively high ability.

Conclusion and Summary

Although the Rasch model apparently can potentially provide the population- and item-invariant scaling needed for objective measurement, it is certainly no panacea for the test developer's problems. Some of the claimed advantages of Rasch scaling 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 assumptions must be met.

The most direct influence of item characteristics is on the population-invariance of item calibrations. The Rasch item parameter estimates will be invariant only under a special condition. Individuals with the same raw score must have the same probabilities of passing each item, regardless of the population to which they belong. Thus, item parameters will not be population-invariant when there is cultural bias which differentially affects the item probabilities. 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. Compared with the classical model, however, the Rasch model is superior since difficulty level is never population-invariant.

Although population-invariance of ability estimates is probably attainable for any item pool, how much of an advantage this is depends on the theoretical interpretability of the item pool. The Rasch ability parameter estimates are nearly as invariant as percentage correct scores, but have the same disadvantage. Interpretation is possible only relative to the existing set of items, the calibrated item pool. As with domain-referenced testing, the items in the set must have a priori validity. In general, the current explication of most trait constructs does not even approach the kind of precision required for a domain-referenced interpretation. Again, the Rasch model offers a potentiality, but does not solve basic theoretical problems in test interpretability.

The major focus of this paper has been on the construction of equivalent forms from a calibrated item pool. 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 and Panchapakesan, 1969). To understand the characteristics of the Rasch model in developing equivalent forms, it was found necessary to distinguish between statistical equivalency, in the narrow sense, and maximal equivalency. Item subsets are statistically equivalent if the differences obtained on some sample are distributed as would be expected from the measurement error associated with each score. Maximal equivalency, however, means that the measurement differences between tests is as small as possible. It was pointed out that using any subset from an item pool calibrated by the Rasch model would lead to statistical equivalency but not necessarily maximal equivalency between subsets.

The empirical results generally substantiated this 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. None of the subsets were maximally equivalent and precision might have been increased by using more efficient techniques in selecting items. The classical techniques of having item difficulties close to .50 for the population and matching extreme item difficulties would then apply if the tests are to be equally precise at each score level.

It may be wondered, then, what advantages the Rasch model really offers, if maximally equivalent forms necessitate using classical item selection criteria. The real strength of the special statistical equivalency of Rasch-calibrated item subsets is the possibility of individualized selection of items rather than the construction of fixed content tests. The unusual characteristics of Rasch measurement errors allow the desired degree of precision for any person to be obtained from the fewest possible items. Estimates of ability and measurement error associated with each possible raw score for any subset of items can easily be determined. 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, will be as close to the ability estimate as possible and will give the largest increase in precision. Tests developed according to classical techniques are not suitable for individualized item selection since measurement error can only be estimated for a whole test actually administered to some population.

In conclusion, the lack of impact of the Rasch model is due more to the current status of trait measurement than to the features of the model. The true advantages of the Rasch model necessitate a more sophisticated technology in trait measurement than is now characteristic of the field. Explicit trait-item theory, culturally-fair items and computer administration of tests would be part of the necessary technological sophistication.

References

Anderson, J., Kearney, G. E., and 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.

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

Popham, W. J. and 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. An item analysis which takes individual differences into account. British Journal of Mathematical and Statistical Psychology, 1966a, 19, 49-57.

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, 1966b, Pp. 89-108.

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., and Panchapakesan, N. A procedure for sample-free item analysis. _Educational and Psychological Measurement_, 1969, _29_, 23-48.

Footnotes

1. The cell values actually used in computations are not the simple likelihoods. A correction is made to prevent infinite values from occurring when all members of the score group pass the item. 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

w = percentage of total calibrating sample obtaining score j

2. This interpretation is oversimplified to maintain conceptual clarity. The actual cell values used in computation are corrected cell likelihoods.

Table 1

Item-by-Score-Group Probability Matrix

Total-Score Group
(Raw Score)

Item $i=1,k$	1	2	3	4	.	.	.	.	$k-1$ ($j=1,k-1$)
1	P_{ij}	P_{i2}	.	.	.	.	.	.	P_{ik-1}
2	P_{2j}								
3									
.									
.									
.									
.									
.									
k	P_{2k}								

Table 2

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

Subset	Raw Score				Log Likelihood				
	$\bar{x}$	$t_{\bar{x}_1-\bar{x}_2}$	s^2	$t\sigma^2$	$\bar{x}$	$t_{\bar{x}_1-\bar{x}_2}$	s^2	$t\sigma^2$	SE msmt.
Odd	19.38		23.67		.432		.873		.453
		17.78		2.99		.86		.51	
Even	15.47		26.85		.415		.892		.433
Easy	22.31		29.43		.469		1.178		.503
		42.52		6.23		2.15		13.76	
Hard	12.53		22.18		.415		.663		.419
Random Set I	17.79		24.25		.427		.882		.447
		2.86		.69		.43		.67	
Random Set II	17.06		25.83		.436		.857		.433

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, Even	.425	.76	.007	1.028	1.014
Easy, Hard	.590	.76	-.057	1.313	1.146
Random Sets	.410	.76	-.020	.995	.998