The Ethiopian Journal of Education, Vol. XIV No. 2, 1993

THE RASCH METHOD OF ITEM ANALYSIS ADVANCES IN ITEM AND TEST DEVELOPMENT

Ayele Meshesha *

ABSTRACT: *The purpose of this article is to recast the Rasch method of item and test development in both theoretical and practical forms to broaden insights into the model's applicability in everyday work and research. The Rasch model tries to describe what happens when an examinee with a certain ability encounters a question with a certain difficulty level. Following some assumptions, it states in a probabilistic manner the way people and questions relate. The process contrasts sharply with the most widely used item analysis techniques, which simply report total scores on people and on questions and item-total score correlation coefficient as the means of describing test response data. The Rasch method uses two pieces of information: item and person parameters. With the help of the computer these parameters are calibrated and indexed. From this a scale free from a particular population of students will be created which will enable the user to interpret level of performance directly with respect to the curriculum.*

* Assistant Professor, Department of Educational Psychology, Addis Ababa University

INTRODUCTION

Systematic approaches to test development with the purpose of getting an accurate measurement of a person's ability have been a prime concern of psychometricians. The quality of items in a test will determine how accurately an ability or trait will be measured. Information about the strength and quality of items in a test can be judged by applying item analysis procedures. To date, the majority of professionals and teachers have been using classical test theory as their basis for development and item analysis.

The classical test model is a weak model and as a consequence has limitations. For example, the classical test theory is sample dependent. Ability parameters that represent the position of the examinee on the variable to be measured depend upon the total group of individuals who take the test.

Classical measurement concepts like reliability and validity coefficients are sample dependent. The former is not applicable to an individual, but only to a group of examinees because the coefficient of correlation involves variation among scores of different examinees.

The value of item difficulty index, P, which is the proportion of correct responses to items in a test, is also tied to the performance of the sample used and varies significantly with different samples. Likewise, the item-test correlation is sample dependent.

The classical test model is not only sample dependent, but is also item dependent. An obtained score depends upon the specific items chosen for the test and reflects the difficulty of the items. One has to give the same set of items to another group of interest to get a precise comparison (Hambelton & Cook, 1977). It is evident, however, that if we fail to get adequate information about ability and item statistics other than for the group to which the test was given initially, we will not be able to derive comparable scores from different tests given to different people. It is clear that test developers, be they professionals or ordinary classroom teachers, will benefit from any procedure that will provide them with comparable score and interpretive information.

The need for comparable scores becomes desirable in some situations where a course is given to a large number of students in different sections and is taught by different instructors. Marks given to students in each section have to be comparable

overall in spite of the difference between teachers. But without a common scale it is difficult for educators to appraise educational achievement from year to year and compare groups from section to section or from school to school.

The weaknesses of the conventional test theory have led experts in the field to seek alternative ways of item selection where the test statistic would be stable even with changes in group ability. As early as 1950, Gulliksen (1950, 392) made a remarkable plea which is worth recalling: "A significant contribution to item analysis theory would be the discovery of item parameters that remained relatively stable as the item analysis group changed or the discovery of a law relating the changes in item parameters to changes in the group".

Many attempts have been made by various people to solve the problem posed by Gulliksen by introducing theories or models for test scores. Among these models are strong true-score and latent trait theories. Even though all models specify relationships between observable examinee performance and an unobservable trait assumed to underlie performance of the examinee, each model has its own special assumptions which make some of them more complex than others.

There is one particular latent trait model, generally referred to as the Rasch model, or the One-Parameter Logistic model, which has been widely tried in item selection and test development. This paper attempts to set down the main features of the method in such a way that it can be understood and can be used in everyday work and research.

Theoretical Considerations in Estimation of Ability and Item Parameters.

In the early 1960s George Rasch discussed a new idea of looking at test data. He based his discussion on a theoretical approach that had risen out of his application of mathematics to statistical aspects of testing. However, it was since the publication of Statistical Theories of Mental Test Scores (1968) by Lord and Novick that considerable attention has been given to the field of latent trait theory in general and the Rasch model in particular as a new area of test development.

Proponents of the new model claim that its advantages over classical test theory are twofold: (1) theoretically it provides item parameters that are invariant across examinee samples that will differ with respect to the latent trait, and (2) it provides

information about how a specific item discriminates among students of varying abilities (Lord and Novick, 1968; Wright, 1977).

The Rasch method proposes a way of understanding and describing test data different from the conventional or classical test development procedures. In classical item analysis techniques the total number of correct responses a person gets on a test determines his rank (ability) in the group and the number of correct responses to an individual item expressed as proportion to the total decides the level of difficulty of the item.

To carry out item analysis in the Rasch method we also employ the number correct for the questions and for the persons taking a test. However, in this model there is a theoretical consideration which describes the stochastic outcome of the interaction of a person with a certain ability level and an item with a given level of difficulty. The individual examinee may get an item right or wrong, but the probability of a correct response depends on both the ability of the person and the difficulty of the item. It can be said, however, that an examinee is more likely to get the easier questions right.

The Rasch model starts at this point. The phrase 'more likely' leads us directly to the use of a probability scale. Two measures have to be dealt with to set up the scale: the difficulty level of a particular test item and the ability level of a student. From these, a mathematical function gives the probability of success.

$$P_{vi} = e^{(\beta_v - \delta_i)} / [1 + e^{(\beta_v - \delta_i)}] \qquad (1)$$

Where

P_{vi} = the probability for an examinee v of answering item i correctly

β_v = the ability parameter for examinee v

δ_i = item difficulty parameter

e = the base of the system of natural logarithms

When a person encounters an item, the result is never going to be clear. An able examinee might miss an easy item while a less able examinee might respond correctly to an item that has been reckoned hard. But the response model states that an able person is more likely to get a question right than a less able candidate.

The person-item relationship shown by their difference $(\beta_v-\delta_i)$ above is used to produce a probability statement for the model. Since either parameter can vary from minus infinity to plus infinity, so can their difference. But probability must stay between zero and one. To overcome this problem, the difference $(\beta_v-\delta_i)$ has to be expressed as an exponent of the natural constant e (Wright & Stone, 1979).

The table given below presents a numerical example to show the way in which people and questions relate. In other words, it shows how the probability P_{vi} works as a function of the difference between person ability and item difficulty.

Table 1. The Rasch Probability of a Right Answer as a Function of Person Ability and Item Difficulty*

Person ability β_v	Item difficulty δ_i	Difference $\beta_v-\delta_i$	Odds $e^{(\beta_v-\delta_i)}$	Right answer probability
5	0	5	148.41	0.99
4	0	4	54.60	0.98
3	0	3	20.09	0.95
2	0	2	7.39	0.88
1	0	1	2.72	0.73
0	0	0	1	0.50
0	1	-1	0.37	0.27
0	2	-2	0.14	0.12
0	3	-3	0.05	0.05
0	4	-4	0.02	0.02
0	5	-5	0.01	0.01

* Partly from "Solving Measurement Problems with the Rasch Model" by Benjamin D. Wright, *Journal of Educational Measurement*, 1977, 14.

One can see from Table 1 that when person ability is more than the difficulty of an item, then, β_v is more than δ_i. The

difference between the two variables is positive and P_{vi} on the item is greater than .50. But when the item is too difficult we find that β_v is less than δ_i and their difference is negative which makes P_{vi} less than .50. As an item becomes more difficult for a person, we see that this probability of success gets increasingly lower, i.e. the probability of success of an individual which is .27 when β_v is 0 and δ_i is 1 becomes .01 when difficulty (δ_i) becomes 5.

In short, this is how examinee ability and item difficulty relate when people take test questions that **confirm** to the specifications of the Rasch model.

Estimation of Parameters

Usually computer programs are used to generate the best parameter estimate for the Rasch model. As indicated later in the "Application of the model" section of this paper, BICAL is a highly portable FORTRAN estimation program for the Rasch model in addition to a few others.

However, there are many different approaches, to the problem of estimation of parameters to the model, which vary in their mathematical and statistical sophistication. For instance, there are simple procedures which can be carried out by the use of hand calculator (Cohen, 1976) which are based upon the assumption of normality of the distribution of examinee ability. This method had been used by Benjamin Wright (1977) and

has been adopted here for the estimation of parameters from a hypothetical example as follows:-

1. For a test of L items given to a sample of N persons; delete all items that no one gets right or no one gets wrong and all persons with no items right or no items wrong and continue deleting until no such items or persons remain. For the L items and N persons remaining:

2. Observe S_i the number of persons who got item i right, for i=1 through L and n_r the number of persons who got r items right, for r=1 through L-1.

3. Calculate

$X_i = \ln [(N - s_i) / s_i]$ the log ratio of wrong to right answers to item i by N persons, (2)

$\bar{X} = \sum_{i}^{L} X_i / L$ the mean of X_i over L items, (3)

$U = \sum (X_i - \bar{x})^2 / (L - 1)$ the variance of X_i over L items, (4)

$y_r = \ln [r/(L-r)]$ the log ratio of right to wrong answers on L items, (5)

$\bar{Y} = \sum_{r}^{L-1} n_r y_r / N$ the mean of Y_r over N persons, (6)

$$V = \sum_{r}^{L-1} n_r(y_r - \bar{Y})^2 / (N-1)$$

the variance of Y_r over N persons, (7)

4. Let $X = \left[\frac{1 + U/2.89}{1 - UV/8.35} \right]^{1/2}$

an expansion factor due to variation in item difficulty, (8)

$$Y = \left[\frac{1 + V/2.89}{1 - UV/8.35} \right]^{1/2}$$

an expression factor due to variation in person ability, (9)

5. Then

$$d_i = Y(X_i - \bar{X})$$

the difficulty estimate of i, (10)

$SE(d_i) = Y[N/s_i (N-s_i)]^{1/2}$

the standard error of difficulty calibration (11)

$b_r = Xy_r$

the ability estimate implied by score r, (12)

$SE(b_r) = X[L/r(L-r)]^{1/2}$

the standard error of ability measurement (13)

As an example of this procedure, suppose 540 persons took a five item test with responses as shown under s_i and n_r in Table 2. Calculation of U, V, X and Y produce the d_i and b_r values listed in the table. Since the data were generated by simulating the exposure of randomly selected persons with mean ability zero and standard deviation of 0.5 to five items with the difficulties shown under δ_i the success of the calibration can be judged by comparing the estimated d_i with the corresponding δ_i values.

Table 2: An Example of Rasch Model Calibration.

Item	S_i	X_i	$d_i = Y(X_i - \bar{X})$	SE(d_i)	δ_i
1	390	-0.96	-1.02	0.11	-1.00
2	357	-0.67	-0.69	0.10	-0.50
3	285	-0.11	-0.05	0.10	0.00
4	202	0.51	0.67	0.10	0.50
5	146	0.99	1.21	0.11	1.00

N = 540 $\bar{X}$ = -0.07 U = 0.653 X = 1.14

Score	n_r	y_r	$b_r = Xy_r$	SE(b_r)
1	83	-1.39	-1.58	1.27
2	170	-0.41	-0.47	1.04
3	197	0.41	0.47	1.04
4	90	1.39	1.58	1.27

N = 540 $\bar{Y}$ = 0.04 V = 0.733 Y = 1.15

Illustrative Example

In order to illustrate the way this technique operates, an example of a simple arithmetic test is given below.

1. 7 + 8
2. 42 - 27
3. 9 x 17
4. 216 ÷ 9
5. 30 ÷ 0.5

In an achievement test whose main function is to distinguish different levels of achievement, an item answered correctly by all examinees, or incorrectly by all, does not help to distinguish between high ability and low ability students. Such items will be excluded from the analysis. Likewise, students getting zero or 100 percent correct marks are eliminated. Anyone who gets all the items right has undoubtedly a high ability, but it is difficult to assign him a specific ability level. Also.

someone who gets all items wrong can be classified in the low ability category but one cannot say how low he or she is.

Let's now examine possible results of 10 students on the five-item test mentioned earlier.

Table 3: Results of 10 Students on Five Questions.

Distribution of Items						
Examinee	1	2	3	4	5	Totals
Askale	✓	✓	✓	✓	X	4
Bogale	✓	✓	✓	X	X	3
Chaltu	✓	✓	✓	X	X	3
Debela	✓	✓	X	✓	X	3
Enanu	✓	X	X	✓	✓	3
Fikre	✓	✓	X	X	X	2
Gonite	✓	✓	X	X	X	2
Hirut	✓	X	✓	X	X	2
Jigsa	X	✓	✓	X	X	2
Konjit	✓	X	X	X	X	1
Total correct	9	7	5	3	1	

✓ = Correct X = Wrong

For illustrative purposes the table has been arranged in a hierarchical manner. The items are arranged in order of increasing difficulty from left to right and the examinees are put in order of decreasing total marks from top to bottom of the table. This is to say that item number 1 is the easiest while item number 5 is the hardest. In like manner, candidate Askale with a total mark of 4 is at the top of the ability list while Konjit is at the bottom.

Examining the table one realizes that it follows a general pattern indicating a rough relationship between item difficulty and examinee ability. Such would be the manner when items and students fit the Rasch model in a real situation. However, for a test of considerable length and a large number of examinees, the pattern will not be as easily observable as in the current assumed example. Usually computer programs are used to generate the best fitting parameter estimates for the model. Here we simply try to illustrate some general principles.

Now let's look at the table again to examine the goodness-of-fit between items and examinees. Considering the students, we see that Askale, Bogale and Chaltu are consistent in the manner of their answers to the questions. Similarly we find that Fikre, Gonite and Konjit are also compatible in the way they reply to the items. None of them scores correct on a more difficult

question than any they get wrong. Debela is some what consistent, although he solves question 4 after missing the easier item number 3. We observe a similar condition in the answers of Hirut and Jigsa where a difficult item has been tackled after getting an easier item wrong. It is Enanu whose results are rather inconsistent; except the first question, she has succeeded in answering the more difficult questions 4 and 5 than the easier items 2 and 3.

There have been several suggestions in the literature for testing the goodness-of-fit of the Rasch Model. The most commonly used testing statistics are those sometimes known as global test statistics. The approach to these is to calculate the probability of all the possible response patterns and then compare the observed and the theoretically expected outcomes by means of a chi square (χ^2) test (Wright & Panchapakesan, 1969; Andersen, 1973; Harris *et al*, 1988). In these statistical tests, the sample under consideration is partitioned into ability groups and the equality of the item parameters over ability groups is checked directly.

To analyze the likelihood of the results scored by the examinees in our example, we will consider Bogale and Enanu. For candidates of the ability level required to get 3 out of 5, the chances of success in the five questions respectively are 0.9, 0.7, 0.5, 0.3 and 0.1. The probability that such a student

gets Bogale's pattern of results is 0.19845 which is the product of 0.9, 0.7, 0.5, 0.7 and 0.9.

For Enanu, by contrast, the outcome is the product of 0.9, 0.5, 0.3 and 0.1, which is 0.00405. Likewise, we can compute the probabilities for the rest of the group. For Debela, the figure comes to 0.08505. We see that Bogale's result pattern is 29 times as likely to occur as Enanu's and even Debela, who has succeeded in answering item 4, has a chance 21 times as likely as Enanu.

Enanu has produced a result so noticeably different from the pattern of results as a whole that it may have to be decided that she does not fit the model and therefore be eliminated. Such would be the procedure followed for testing the goodness-of-fit for longer tests involving a large number of examinees.

In the same way as for the students, the selection of the items best fitting the model is necessary. Checking the table, we find that none of them seriously diverges from the expected pattern. In an actual test, the elimination of items is of rather greater consequence than the elimination of students, but the method of analysis to be used will be identical to that applied in the example involving Enanu and her friends.

After the elimination process is completed, we will be left with a group of items that fit the model. Each question or item in the group will have a calibrated difficulty level. More items can be calibrated in this manner and can be stored in the computer to form what is known as a bank. Then when it becomes necessary, a part of the collection may be used to test a group of students whose achievement can be compared with that of the original group.

Once the bank has been formed, a test or several tests can be made according to desired difficulty levels from the already known difficulty level of each individual item. For students who do not fit the model their results could be used to analyze areas of weakness for future remedial actions.

To develop a viable system that can accommodate the ideas discussed above, the various assumptions about the model have to be explored

Assumptions for testing Goodness-of-fit

All latent trait models are based upon a set of basic assumptions as to what happens when people take a test. The Rasch model assumes that the items are measuring one common ability and there exists the assumption of local independence

between the items and the examinees. These two assumptions imply that a test which measures only one trait or ability will have less measurement error in the test score than a test that is multidimensional, and that the response of an examinee to one item is not related to his response to any other item. The model also makes a third assumption, that all items have equal discriminating power but vary only in difficulty. There is an additional assumption that the items cannot be answered correctly by guessing.

Wright (1977) has suggested that the Rasch model should be superior to other latent trait test development procedures due to the simplicity of the model. That is, the unweighted number right scoring technique used by the Rasch model contains all the information necessary to produce estimators for item and person parameters. Considering the model's desirable features, Hulin *et al* (1983:38) have this to say:

> This model is perhaps most useful when a researcher has carefully pretested a set of items that were written in a format that minimizes guessing. Then it may be possible to select a subset of these items with approximately equal discriminating powers. Under these conditions the simplicity of the one-parameter model makes it very attractive to practioners.

Lord (1980) assures us that if the assumptions are satisfied for a set of data, sufficient statistics are available for estimating both item difficulty and examining ability. He also argues that if sample size is small, Rasch estimates may be more accurate than the three-parameter-model estimates.

We can take some of these assumptions and consider briefly the conditions in which they might be justified.

The first is that all questions must have levels of difficulty which may be compared directly with each other. Taking the illustrative example given earlier, if question 4 on division proves to be more difficult than the multiplication question (3), then it is assumed that this is so for all students. If there are some students who find division generally easier than multiplication, while others find multiplication easier, then these two areas must be tested separately.

In other words, there must be a common scale of measurement for the difficulty level of all items. The test quoted might be said to measure ability in simple computation, and its use presupposes the existence of a single scale of measurement for this purpose. If the scales of measurement for ability in say, multiplication and division are different, then the model cannot be used for a test which includes items on both. To take an extreme example of this, one could not use a test consisting

partly of computation questions and partly of questions on English grammar, because the two sets of questions are not measuring the same thing and so the scales of measurement are different. The fundamental requirement is that all students should respond to all **items** in a similar manner, even though the levels both of difficulty and ability may differ.

In practice, it is not obvious whether a given set of questions is measuring just one thing, or two, or more than two. In many cases an initial judgement would be that there were probably several, but that they were sufficiently alike to be considered together without departing significantly from the assumption of a single scale. No firm conclusions can be drawn without setting up a full run of the test, with a large number of students, and looking at the results.

If it were necessary to observe the requirement of a single scale very strictly, the possible use of the Rasch model would be very limited. But it is frequently stated by the proponents of the method that minor variations from the single scale are acceptable: in other words, all items in a test must measure at least roughly the same thing. The real nature of the assumption is thus that the variations which exist are small enough not to cause significant errors.

The second assumption is that the level of difficulty of a particular question is independent of its context: that it does not matter whether it comes first or last or whether it is preceded by easier questions of the same type. In other words, it is assumed that a particular set of questions may be arranged in any order, without affecting the results.

In many tests it is not unusual to arrange the order of questions for the student's benefit. One might start off with a few very easy items to give the weaker students confidence, then have several graded groups of items on different aspects of the subject, and leave a few really difficult questions until the end. Often a question will be set in several parts so that success in the first part is necessary before the remaining parts can be attempted. When using the Rasch model, this cannot be done without departing from the conditions under which it operates.

Nevertheless some questions have to come at the start and others at the end. The assumption, that this makes no difference, is a big one because it is beyond dispute that students do learn during a test. They learn through practice; they learn to do difficult items by doing easy ones. They may pick up hints about the meanings of words or other information which can be used in questions other than the one intended.

The real point at issue is not whether local independence does in fact exist. It is whether the actual degree of interdependence is great enough to cause significant error.

The third assumption is that there is no guesswork effect. Given the essential requirement that questions must be dichotomously scored, it is possible to use either questions to which the student has to provide an answer, or questions in which he has to choose between several answers given on a test paper. Also, a single answer may be given whose correctness or otherwise the student has to decide. Clearly there is scope for guesswork, particularly when possible answers are given, and most of all in the true or false type.

It is not necessary to assume that there is no guessing at all, but rather that the amount of guessing is so little and in such a pattern that the operation of the model is not significantly affected.

In general, it can be concluded that if the conditions can be met, and if the errors from all sources do not become too great, we have a method of analysis of a considerable power.

Thus far the discussion has focused on the necessary specifications and theoretical background of the Rasch method to

understand how it works. The following section will deal with applications of the model in item and test development.

Applications of the Rasch Model

Much has been written about the use of the Rasch model to real world simulations. Within the last decade or so many researchers have been interested in the empirical investigation of the model's robustness with respect to some aberrations like guessing and unequal item discrimination slopes. Some have reported that the model is extremely robust under many situations while a few others have found less encouraging results.

In a resent study Forsyth *et al* (1981) examined the invariance properties of the Rasch model using data that did not confirm in all respects to assumptions of the model. They used sections of the Iowa Tests of Educational Development which were built according to content-by-process table of specifications. They found that the Rasch model yielded reasonably invariant item parameters and ability estimates even though the assumptions of the model were violated.

The Rasch Calibration method has also been applied to a number of school-related content areas and the analysis has

been reported successful. Soriyan (1972) used the model to find out to what degree achievement tests of the West African Certificate Examination would behave in terms of fitting the model. He found that the test satisfied the assumption of constant item discrimination. It was also reported that the model did well in selecting items even in those instances where discrimination indices were unequal.

Some investigators have raised the issue of adequate sample size to be used in the Rasch model. It has been found that sample size does not matter at all. Tinsley and Davis (1975) report that the Rasch item difficulty index and estimation of ability were invariant. They used as small as 25 items in some of their tests and samples of 89, 120 and 145 examinees. The largest group sample they tested consisted of 630 students.

The examination of the goodness-of-fit of the items to the Rasch model and the processing (item calibration) of the data are carried out by the computer. There is a host of programs available for both mainframe and personal computer use (Mislevy & Stocking, 1989; Mislevy & Bock, 1986). However, the most often used software for the estimation of parameters and calibration of items has been the BICAL program. It implements a maximum likelihood estimation procedure for concomitantly estimating item and person parameters from observed item responses. The program

anchors the location of the difficulty and ability scale metric by setting the mean value at zero (Haberman, 1977; Van den Wollenberg *et al*, 1988).

The point (zero) reference for ability levels has been difficult to explain to students and parents. A handy translation and scaling to units called "WITs", provides units which can be expressed in terms of positive integers. The transformation is $d = (100+9.1\delta)$ and $b = (100+9.1\beta)$. Other scale names like "logit" and "RITs" are sometimes used to indicate ability levels of examinees.

It is necessary to note that the test analysis process in the Rasch model contrasts sharply with the most widely used item analysis techniques which simply report total scores on people and on questions and item total score correlation coefficients as the means of describing the collected data. The Rasch model, however, enables us to examine and understand the nature of the test responses collected from individuals rather than just to describe them.

The standard error of measurement calculated from the internal consistency reliability estimates is the same for any score obtained on the test as a whole. With the Rasch method it can be shown that at the centre of the distribution of scores, roughly around the mean, the standard error of attainment is

approximately equal to that obtained from using the classical item analysis approach. However, as the Rasch estimates of pupils' attainment become more extreme from around the mean, the standard errors derived from using the model do increase, indicating the rather less precision with which the estimates are obtained. This is an advantage one gets from the new method and saves test users from assuming that the standard error of measurement would be the same over the whole range of test scores.

The Rasch analysis is extremely useful in that area of test development known as item bank construction and management where large scale calibration exercise can be done. The initial procedure for establishing an item bank is somewhat complex and time-consuming. Thousands of questions which are appropriate and educationally meaningful have to be collected and administered over several sessions in order to calibrate a substantial number of relevant items into an item bank. The complexities are a result of not only collecting appropriate data but also of the explicit assumptions and the need to check for the compliance of such assumptions.

But once an item bank has been established with the necessary item statistics and other item characteristics, one will be able to select items for a test on the bases of that information, as well as its content. Besides, one will be able to check from

time to time if the bases for using the model are met every time data are collected and items are used from the bank. If the conditions are not met, then the proper examination will be carried out to check in what respect and to what degree conditions have been violated.

It is here that we get the main difference between the use of the Rasch model and the techniques based on classical item analysis. In most classical item analysis methods, items are simply clustered together to form either sub-scores or a single total score. It is not a common practice to check in detail to ensure that data from a particular examination session are indeed measuring something which is consistent with what has been the case in the past as measured by the individual items. As stated earlier, in the Rasch method, however, we will know if there are major departures from the model and do something about them.

The use of the Rasch method becomes apparent, especially, for large scale testers which have the necessary manpower and equipment. The introduction of an item bank, its management and use would be facilitated by the use of this new tool. Once after we have included all the items we want, we can use them in any order we desire. Or we can create as many alternate forms of a test as we need and be able to administer it to a

large body of candidates without much worrying about cheating or copying by examinees.

In addition, as indicated in the introduction part (pages 1 & 2), the use of the same test to assure comparability of information is most often necessary. The same result can now be achieved without the problems of test security and out-datedness of test content by using an underlying curriculum scale and a Rasch-calibrated item bank.

Conclusion

It has been demonstrated extensively both in research settings and practical applications that the Rasch method is useful for item banking and test development. In other words, the model can be used to define any described variable and to develop, field test and calibrate corresponding test items.

Items that qualify are stored in the bank with hierarchical levels of difficulty. They can be retrieved and tests easily created from them in several parallel forms as needs arise. Such systems will definitely facilitate the work of many educational and industrial institutions that occasionally use tests.

From the technical point of view, much has been done in the area of computer technology over the last decade or so. To successfully develop, field test and calibrate items of any size, we had to use mainframe computers which were very expensive to acquire. That has now changed significantly.

The technological advances made in the computer world where documents containing text and graphic can be published on desktop machines, maximize the capability and possibility of introducing computerized testing even to the classroom level. On this point of applying computers for testing packages, Frank Baker (1990:18) says:

> A few years ago, we were talking about how nice it would be to administer tests via the microcomputer and now people are doing it on a rather large scale. Much of this advance rests upon the dramatic increase of computer power available at a reasonable cost and upon the existence of commercial testing software. Because of the latter, we are beginning to get away from reinventing the wheel each time a new testing situation arises.

The commercial testing packages that Baker is talking about are now available in the market. Organizations like Assessment Systems Corporation of St. Paul Minnesota produce computerized testing products such as MicroCAT (testing

system), ASTEC (Exam creation system) and ITEMAN, RASCAL and ASCAL (Item analysis programs). The complete package will be a few thousand U.S. dollars (Patience, 1990).

It is understandable that financial resources must be available to hire competent personnel to adapt the software, to develop and maintain item banks. But these days costs are not that prohibitive and when considering the advantages and the impact the system will have on our education, it will be money well spent.

REFERENCES

Andersen, E.A. (1973). "Goodness-of-fit Test for the Rasch Model". Psychometrika, **38**:123-140.

Assessment Systems Corporation (1990). Computerized Testing Products Catalog. St. Paul, mn: Assessment Systems Corporation.

Baker, Frank B. (1990). "Some Issues in the Application o f Microcomputerized Testing Packages." In Educational Measurements: Issues and Practices.

Cohen, L. (1976). "Approximate expression for parameter estimator in the Rasch model." British Journal of Mathematical and Statistical Psychology 32:113-120.

Forsyth, R., Saiangjan, U. and Gilmer, J. (1981). "Some empirical results related to the robustness of the Ranch Model". Applied Psychological Measurement 5:175-186.

Gulliksen, H. (1950). Theory of Mental Tests. Wile.

Haberman, S. (1977). "Maximum likelihood estimates in exponential response models". The Annals of Statistics 5:815-841.

Hambelton, R. and Cook, L. (1977). "Latent trait models and their use in the analysis of educational test data". Journal of Educational Measurement 14:75-96.

Harris, J., Laan, S. and Mossenson, L. (1988). "Applying partial credit analysis to the construction of narrative writing tests". Applied Measurement in Education 1:335-346.

Hulin, Chanles l. *et al.* (1983). Item Response Theory: Application to Psychological Measurement. Homewood: Dow Iones Irwin.

Lord. F.M. (1980). Applications of Item Response Theory to Practical Testing Problems. Hillsdale, NJ: Erlbaum.

Lord, F.M. and Novick, M. (1968). Statistical Theories of Mental Test Scores. Reading, Ma. Addison-Wesley.

Mislevy, R. and Bock, R. PC-BILOG. (1986). "Item Analysis and Test Scoring with Binary Logistic Models (Computer Program). Mooreville In Scientific Software.

Mislevy, R. and Stocking, M.A. (1989). "Consumer's guide to LOGIST AND BILOG". Applied Psychological Measurement **13**:57-75.

Patience, W. (1990). "Software Review". Journal of Educational Measurement **27**:82-88.

Soriyan, M. (1972). "Measurement of the goodness-of-fit of Rasch's probability model of item alysis to objective achievement tests of the West African School Certificate Examination". (Doctorial Dissertation, University of Pittsburg 1971). Dissertation Abstract International **32:** 4433A.

Tinsley, H. and Davis, R. (1975). "An investigation of the Rasch simple logistic model: sample free item and test

calibration". Educational and Psychological Measurement **35:** 325-339.

Van den Wollenberg, A., Wierda, F. and Jansen, P. (1988). "Consistency of Rasch model parameter estimation: A simulation study". Applied Psychological Measurement **12**:307-313.

Wright, B. (1977). "Solving measurement problem with the Rasch Model". Journal of Educational Measurement **14**:97-116.

Wright, B. and Panchapakesan, N. (1969). "A procedure for sample-free item analysis". Educational and Psychological Measurement **9**:23-48.

Wright, B. and Stone, M. (1979). Best Design, Rasch Measurement, Chicago: MESA Press.

THE ETHIOPIAN JOURNAL OF EDUCATION GUIDELINES TO CONTRIBUTORS

1. GENERAL

1.1 The Ethiopian Journal of Education publishes scholarly articles based on work in education and related areas.

1.2 Besides original research papers, EJE publishes book reviews, dissertation abstracts, short communications and comments on articles published in EJE.

2. STYLE AND FORMAT

Before submitting the manuscripts for publication in EJE, authors should ensure that the following requirements are complied with:

2.1 Title Page

2.1.1 The following shall appear on the Title Page:-

a. the full title of the article;

b. the name(s) of the author(s);

c. the title(s), academic position(s) and affiliation(s) of the author(s) referred to at the bottom of the page

with the use of an asterisk if it is a single author or numerical subscripts against each name.

2.1.2 It is the responsibility of the authors to declare the degree of contribution made by each of them to the preparation of the study. But normally, the following rule shall apply;

a. equal contribution is presumed when the names are written in alphabetical order; or

b. the degree of contribution shall be determined by the order in which the names appear, unless indications are given by the authors to the contrary.

2.13 All correspondences will be made with the author whose name appears first (unless indicated otherwise).

2.2 Length of an Article

2.2.1 Manuscripts should not exceed 30 pages, including an abstract in about 100 words which should be provided on a separate page.

2.2.2 The manuscripts should be typed double spaced on one side of A4 type white paper. A space of one inch should be left on the left and right margins as well as at the top and bottom of each page.

2.3 Citation of Notes and References

2.3.1 All materials, refereed to or quoted must be acknowledged. Plagiarism is illegal and unethical.

2.3.2 Direct quotations should be as short as possible and should be reproduced exactly in all details (spelling, punctuation and paragraphing).

a. short quotations of less than four lines are run into the text and enclosed in quotation marks.

b. long quotations (i.e. more than five lines) should be set off from the text in a separate paragraph, indented (four spaces) and single spaced. Quotation marks are omitted.

2.3.3 References in the text should read as follows:

. Smith (1992:42) has suggested that ...
or

. One educator (Flanders, 1970:16) has argued that...

. Use "*et al*". when citing a work by more than three authors.
Example: Interaction analysis (Flanders *et al.*, 1970) suggests...

The letters a, b, c and so on should be used to distinguish citations of different works by the same author in the same year.
Example: Daniel (1985a, 1985c) recommended that...

2.3.4 Essential notes should be indicated by consecutive subscript numbers in the text and collected on a separate page at the end of the text, titled 'Notes'. Such numbered notes should be kept to a minimum.

Numbered notes should be used to make clarifications about the references used, to include points left out in the text, or to add some items readers may want to know.

2.3.5 All references cited in the text and other supporting materials should be listed alphabetically by author in a section titled References or Bibliography and appearing after Notes. Ethiopian Authors should be listed in the alphabetical order of first name. Daniel Taddesse, for example, should be listed under D and not under T. Write Ethiopian names in full in the Bibliography (i.e. first and second names) as they are given in the publication you are citing. Avoid using honorific titles such as Ato, Dejach, Dr, Wzro, etc. in citation or references.

A. Published Articles

The following are examples of different entries in References or Bibliographies.

i) Kremmer, L. (1978). "Teacher's Attitude Towards Educational Goals as Reflected in Classroom Behaviour", Journal of Educational Psychology, 70,6: 993-997.

ii) Ayalew Shibeshi (1989). "Some Trends in Regional Disparities in Primary School Participation in Ethiopia", The Ethiopian Journal of Education, X,1: 25-51.

Note: The volume and issue numbers should be entered exactly as they are given in the journals cited (i.e. in Roman or Arabic numerals).

B. Books

i) Perrott, E. (1982). Effective Teaching: A Practical Guide to Improve Your Teaching. New York: Longman Inc.

Listing of several works by the same author should be in the chronological order of the year of publication. Here is an example:

ii) Ryans, D.G. (1989). Characteristics of Teachers. New Delhi: Starling Publishers(p) Ltd.

iii) _____ (1972). Analysing Teaching. New York: Macmillan Co. Ltd.

C. Contributions in Books

Philip, W.J. (1986). "Life in Classrooms" in Norris G. Haring, Analysis and Modification of Classroom Behaviour, pp. 13-17. New Jersey: Parentice-Hall, Inc. 1972.

D. Contributions in Proceedings

Marew Zewdie and Fanta Suppa, Attitudes of Teachers Towards the ESLCE. In Proceedings of the Workshop on Major Issues Related to the ESLCE and Possible Solutions, Nazareth 25-27 April 1991, pp. 235-257, Addis Ababa, Institute of Educational Research.

E. Conference/Seminar Papers

Amare Asgedom (1990). Communication Theories and Instructional Practice: A Limited Effect Perspective; Paper presented at the First

Annual Seminar of the Faculty of Education, 17-20 May, 1990. Nazareth, Ethiopia.

F. Unpublished Works

Tirussew Teferra (1989). The Psychology and Educational Problems of Handicapped Students in Addis Ababa University, A Research Report, Institute of Educational Research, Addis Ababa University.

3. OTHER IMPORTANT RULES TO CONSIDER

3.1 Tables and diagrams:

Tables and diagrams should be properly labelled and carefully drawn. They should have short titles. All footnotes to tables and all sources should be placed under the table.

3.2 Section Headings:

Major section headings must be centered on the page. Sub-headings must be aligned with the left margin.

3.3 Language:

English and Amharic are the Languages of publication. All authors must avoid sexist and racist language.

3.4 Responsibility for Views

Any statements in an article accepted for publication remain the sole responsibility of the author and should in no way be construed as reflecting the opinions of the Editors or the Publisher.

3.5 Copyright

Authors submitting manuscripts do so on the understanding that if they are accepted for publication, copyright to the articles, including the right to reproduce in all forms and media, shall be assigned exclusively to the publisher.

3.6 Originality of the Paper

EJE publishes only original investigations. Authors are not allowed to submit the same manuscript for concurrent consideration by another journal. Already published manuscript should not also be submitted to EJE for publication.

ACKNOWLEDGEMENTS

The Institute of Educational Research is grateful to the continued financial support by SAREC (the Swedish Agency for Research Cooperation with Developing Countries) and to the good office of the Ethiopian Science and Technology Commission.