RASCH ITEM ANALYSIS BY HAND

by

Benjamin D. Wright
University of Chicago

and

Graham A. Douglas
University of Western Australia

Research Memorandum Number 21

STATISTICAL LABORATORY,
DEPARTMENT OF EDUCATION,
THE UNIVERSITY OF CHICAGO

September 1976

RASCH ITEM ANALYSIS BY HAND

Introduction

One of the ironies of progress in psychometrics is that as theories of measurement become more and more sophisticated, the amount of computation required to produce useful results becomes more and more overwhelming. When computers do the work conveniently we may not worry much about the volume of calculation. However, there are situations in which would-be practitioners do not have easy access to a suitably programmed computer and so must either give up or fall back on the needlessly inadequate traditional methods.

Nowhere in psychometrics is this more apparent than in item analysis and test construction. Wishing to be knowledgeable and up-to-date, practitioners turn to latent trait theories of measurement (Rasch, 1960; Lord and Novick, 1968), only to find that in order to determine the estimates of item difficulty, they must be able to solve series of implicit equations which seem quite beyond hand calculation (Andersen, 1972; Lord, 1968).

In 1969 Wright and Panchapakesan described a procedure for item calibration based on the measurement theory of G. Rasch (1960, 1966). Although the procedure still involved considerable computation, the authors justified an approach in which the computations were greatly simplified. In a further development of these ideas, Wright and Douglas (1975b), compared three procedures for use with the Rasch

model from the point of view of practical efficiency and accuracy and gave a simplified account of the Wright-Panchapakesan "unconditional" method (UCON).

In this article we will describe a method of estimation so simple and straightforward that it can easily be done by hand. This method will be referred to as the "normal approximation" method (PROX). We will show that it produces item difficulty estimates (and their corresponding standard errors) which compare favourably with those obtained from the more computationally demanding UCON method. The PROX method relies on a working assumption about the distribution of person ability in the calibrating population and item difficulty in the test, an assumption the utility of which we will document and for which we will detail the effects of departure.

Whilst we offer in the appendix a Fortran program for the PROX calibration algorithm (one which effects a dramatic saving in time and cost over that required by UCON), the simplicity of PROX makes it possible to set up the procedure in a form convenient for hand calculation. First we will discuss the basis of our approximations, then the equations themselves, and finally present results from simulated and real data which support the accuracy of PROX.

The Distributional Assumption

Traditional methods of item analysis which use the

proportion of subjects correct on a given item directly as the estimate of item "difficulty," are sample-dependent. It was towards dealing with this awkward situation that Rasch developed his simple logistic model and for which the term "sample free" test calibration was introduced (Wright, 1968). For the PROX method we use a distributional assumption, not because we believe that abilities are actually distributed normally, but because we have found that the observed ability distributions of most calibrating samples as well as the difficulty composition of most tests can be well enough described by their mean and variance to provide accurate sample free item estimates from these two statistics alone. Indeed, Lord (1955) claims that observed test-score distributions are approximately normal. Sabers and Klausmeier (1971) report that the "skewness and kurtosis indices derived from samples drawn from simulated normal populations varied more from normal values than those observed in a random sample of 200 actual score distributions."

In our work on best test design (Wright and Douglas, 1975a), where we were concerned with achieving optimal measurement in the presence of calibrated items, the crucial question was: what distribution of items in a test achieves the most precise measurement for the most persons? In the course of that study we considered a normal distribution of item difficulties with the bulk of the items "at the

center" and a few scattered at the extremes. A consequence was the derivation of explicit expressions for the person measure, or ability estimate b_r, and its standard error $SE(b_r)$, for any raw score r, when items were normally distributed in difficulty with a known mean and standard deviation.

The derived expressions were found to be exceptionally accurate approximations to the maximum likelihood estimates. Because of the symmetric way in which item and ability parameters enter into the Rasch model, we also investigated the situation in which the population of abilities from which the calibration sample was drawn was also assumed to be normal. The use of this idea to expedite the calibration of items was first suggested by Leslie Cohen in 1972 and has since been further developed by him (Cohen, 1976). We include his most recent suggestions in the basic expressions from which the PROX calibrating algorithm is derived.

The Calibrating Expressions for PROX

If the set of item parameters selected to form a test are distributed normally with mean $\delta.$ and variance $(\gamma\tau)^2$ and administered to a sample with ability variance $(\lambda\sigma)^2$, then explicit expressions for the estimate of the latent ability b_r and its standard error, $SE(b_r)$ corresponding to a raw score r, are given by

$$b_r = \delta. + \lambda \ln\left(\frac{r}{L-r}\right) = \delta. + \lambda b_r^0 \qquad (1)$$

$$\lambda = \left[\frac{1 + \tau^2/2.89}{1 - \sigma^2\tau^2/8.35}\right]^{1/2} \qquad (2)$$

and $$SE(b_r) \simeq (\lambda L/r(L-r)^{1/2} \simeq C_r/L^{1/2} \qquad (3)$$

where (i) there are L items in the test,

(ii) the constants $8.35 = 2.89^2 = 1.7^4$ arise because these approximations use the relation between the normal and logistic cumulative distributions, $|\Phi(x) - \Psi(1.7x)| < .01$ for all x, as a basis for exchanging them in the derivation of the approximation, and

(iii) the error coefficient C_r is for all practical purposes confined between 2 and 3 and varies with the relative score f_r and test width $(\gamma\tau)^2$ as given in Table 3.

In the reverse situation we assume that the population of persons from which the calibration sample has been drawn is normally distributed with mean $\beta.$ and $(\lambda\sigma)^2$. The analogous expressions are

$$d_i = \gamma\left[\log\left(\frac{N-S_i}{S_i}\right) - \frac{\sum_{i=1}^{L} \log\left(\frac{N-S_i}{S_i}\right)}{L}\right] \qquad (4)$$

$$= \gamma d_i^0$$

$$\gamma = \left[\frac{1 + \sigma^2/2.89}{1 - \sigma^2\tau^2/8.35}\right]^{1/2} \qquad (5)$$

$$\text{and } SE(d)_i \simeq [\gamma N/S_i(N-S_i)]^{1/2} \simeq C_i/N^{1/2} \qquad (6)$$

where (i) N is the number of persons in the calibrating sample,

(ii) S_i is the number of persons with item i correct (known as the item score), and

(iii) d_i is the estimate of the item difficulty.

The difference between (1) and (4) is due to the requirement that the item parameter estimates be anchored in some way quite independent of the distribution of abilities; a convenient way to do this is to center them at zero. Equation (4) satisfies this condition and simplifies equation (1) to $b_r = \lambda b_r^0$.

Whilst we may assume the distributional form of abilities to be normal, we must estimate the parameters λ, γ, τ^2 and σ^2 from observed data. In (2) and (5) we will replace τ^2 and σ^2 by their sample statistics,

$$V_d = \sum_{i=1}^{L} (d_i^0)^2/(L-1)$$

$$\text{and } V_b = \sum_{r=1}^{L-1} n_r (b_r^0 - b_\cdot^0)^2/(N-1),$$

$$\text{where } b_\cdot^0 = \sum_{r=1}^{L-1} n_r b_r^0/N,$$

and use them to calculate estimates of the expansion factors,

$$\hat{\gamma} = \left[\frac{1 + V_b/2.89}{1 - V_b V_d/8.35}\right]^{1/2} \qquad (7)$$

and

$$\hat{\lambda} = \left[\frac{1 + V_d/2.89}{1 - V_b V_d/8.35}\right]^{1/2} \qquad (8)$$

A Computing Algorithm

A concise implementation of the above procedure, suitable for computer programming, is facilitated by the following steps.

(i) Edit the binary data matrix of person by item responses such that no person has a zero or a perfect score and no item has a zero or a perfect response. This editing may go beyond a single stage when the removal of an item necessitates the removal of some persons (and vice-versa). The final outcome is a vector of item responses (S_i) and a vector of raw score frequencies, (n_r).

(ii) Let

$$d_i^0 = \log\left(\frac{N-S_i}{S}\right) - \frac{\sum_{i=1}^{L} \log\left(\frac{N-S_i}{S_i}\right)}{L} \qquad i=1,L$$

$$b_r^0 = \log\left(\frac{r}{L-r}\right) \qquad r=1,L-1$$

$$b_{\cdot}^{0} = \sum_{r=1}^{L-1} n_r \, b_r^0 / N$$

$$B = \sum_{r=1}^{L-1} n_r (b_r^0 - b_{\cdot}^{0})^2 / 2.89(N-1)$$

$$D = \sum_{i=1}^{L} (d_i^0)^2 / 2.89(L-1)$$

$$G = BD$$

(iii) Calculate the expansion factors

$$X = [(1 + D)/(1 - G)]^{1/2}$$

$$Y = [(1 + B)/(1 - G)]^{1/2}$$

(iv) Estimate the item difficulties as

$$d_i = Y \, d_i^0 \qquad i = 1,L$$

(v) with standard errors of

$$SE(d_i) = [Y \, N/S_i(N-S_i)]^{1/2} = C_i/N^{1/2}$$

with values for C_i from Table 3.

(vi) The ability estimates for this set of items are given by

$$b_r = X \, b_r^0 \qquad r = 1,L-1$$

(vii) with standard errors of

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

where f = r/L and values of C_f come from Table 3.

A Hand Algorithm

In response to complaints that modern test theory is too complicated and difficult to compete at the practical level with the traditional approach, we describe an algorithm which can be done on a hand calculator. While it can be worked to any desired degree of accuracy, the usual magnitude of the standard errors involved suggests that one decimal place in logits is sufficient. We recommend the short-cut expression for a standard deviation proposed by Mason and Odeh (1968),

$$\hat{\sigma} = 2[\text{Sum Top 6th} - \text{Sum Bottom 6th}]/(N-1)^{*}.$$

Calculations are expedited by a reference table of the relation between proportions to increments of 0.1. Greater accuracy may be achieved by taking smaller increments.

The successive steps in Table 2 have been sequentially numbered and represent the following operations:

(1) Use proportion correct on a given item to write down a frequency distribution of the L items.

(2) Multiply the logit corresponding to each proportion by the frequency for each item and sum the positive and negative totals separately.

---

*When an exact sixth of observations cannot be obtained, proportional parts should be used.

(3) Find the mean of these initial uncentered estimates.

(4) Subtract this mean from each logit to find the initial centered item estimates d_i^0.

(5) Using relative score f = r/L for the persons, write down the frequency distribution of the N persons.

(6) Sum the ability estimates for the top sixth and the bottom sixth persons and estimate the variance of initial abilities, V_b, via Mason and Odeh's short-cut expression, and hence $B = V_b/2.89$.

(7) Sum the item estimates for the top sixth and the bottom sixth items and estimate the variance of difficulties, V_d, via Mason and Odeh's short-cut expression, and hence $D = V_d/2.89$.

(8) Use B and D to calculate G = BD and the expansion factors Y and X.

(9) Multiply each d_i^0 by Y to obtain d_i.

(10) If the final b_f are desired, multiply each b_f^0 by X.

To determine an estimate of the standard error of each item estimate use Table 3 in which an error coefficient for each combination of proportion correct and standard deviation is given. The standard error of item difficulty is this error coefficient divided by the square root of the number of persons taking the test.

An idea of the accuracy of the PROX method (either by

computer or hand) is gained if we compare the estimates in Table 2 with those obtained by the computer algorithm and those obtained from UCON. Table 4 displays all three sets of results. Wright and Douglas (1975a) have studied the degree of imprecision which may be tolerated in item estimation before measurement becomes noticeably biased. In view of those results, namely that one decimal place accuracy in logits is more than enough, we have expressed all three sets of estimates to two decimal places only. Clearly there are no practical differences among the three estimates for any of the items.

Comparison of UCON and PROX for Accuracy--Simulated Data

The simplicity and accuracy of PROX justifies its inclusion as a competitor of UCON. Further justification is forthcoming if we subject the procedure to simulated situations which place increasing stress on the underlying assumptions. In this way we gain information about the degree of departure from normality which may be tolerated before we have to abandon PROX.

In view of the fact that both item and ability distributions are assumed normal, our simulation study challenged both these features. Cases in which the distribution of person abilities is off-center and skewed and item difficulties are both uniformly and normally distributed were

investigated. The simulation was carried out in the following manner:

(i) First a set of 20 and a set of 40 uniformly distributed item difficulties was generated with mean zero and width 4, (e.g., in the case of 40 items, the generated item difficulties would be -1.95(0.1)1.95).

Then a set of 20 and a set of 40 normally distributed item difficulties was generated with mean zero and standard deviation 1.

This produced four separate tests to be calibrated.

(ii) Parameters for twelve different calibrating samples were specified to cover a wide range of possibilities. Emphasis was placed on samples for which the test was off-center (in this case too easy) because these are situations which produce the poorest difficulty estimates. Each sample comprised 500 abilities normally distributed on mean $\beta.$ and standard deviation σ, and restricted by an upper truncation at 5.0 in order to induce skew in the ability distribution. The values of $\beta.$ were 0(1)4 and the values of σ were 0.5, 1.0 and 2.0.

(iii) In any simulation discrepancies between the generating item parameters and their estimates will occur because of the stochastic aspect of the probability model. Long run comparisons between parameters and estimates

are clarified by replication. Instead of administering each of the four tests to a single example of each of the twelve samples, we administered each test to four examples of each sample, each drawn independently from the same normal population. Thus four tests by 12 samples by four replications produced 192 administrations.

(iv) Each "administration" of a test to a calibrating sample was accomplished by simulating stochastic response patterns according to the Rasch logistic response model. The outcomes were the data vectors (S_i) of item scores and (n_r) of person score frequencies which are sufficient for UCON and PROX calibrations.

(v) An editing routine, described in Wright and Douglas (1975b), ensured that these vectors conformed to the algebraic requirements necessary for finite estimates.

(vi) Each pair of data vectors was analyzed by a program which estimates item parameters by means of both UCON and PROX.

The results of these simulations were summarized as follows:

(vii) For each replication of one of the four tests to each of the 12 distributions the UCON and PROX

estimates of each item parameter were calculated, and averaged across the four replications to produce $\bar{U}_i$ and $\bar{P}_i$ and theirabsolute difference $\Delta_i = |\bar{U}_i - \bar{P}_i|$ for each item.

Table 5 illustrates with the statistics from one of these 48 summaries a set of 20 uniform items administered four times to the sample distribution with mean 2 and standard deviation 2. The generating item parameters are also included. Only the results for items 1-5 and 16-20 are shown since, in all cases studied, central items (with $|\delta_i|<1.00$) had Δ_i values of 0.05 or less.

In other work we have found that when L is greater than 20, random values of Δ_i as high as 0.50 have negligible effects on measurement. The maximum observed Δ_i over all cases studied was $\Delta_i = 0.25$ for the easiest item in the 40 item uniform test when administered to a very "smart" sample with mean 3 and standard deviation 1. For this most deviant case the average skewness and kurtosis of person scores were -1.53 and 5.86 respectively--values well removed from the 0.00 and 3.00 expected from a normal distribution. In our summary of all 48 administrations we concentrate on Δ_i in excess of 0.20 and 0.10.

In Table 5 we note that with a mildly skewed distribution, no Δ_i exceeds 0.15. There is no consistent pattern of bias in either the UCON or PROX estimates when compared to

the generating item parameters. The trend in Δ_i is atypical among the 48 administrations.

The results of all simulations are summarized in Table 6. We see that Δ_i's in excess of 0.20 occur only 4 times out of the 1440 item calibrations studied. Δ_i's between 0.10 and 0.20 occur only 30 times. Although an approximately normal distribution of item difficulties is theoretically ideal for the use of PROX, in most calibration situations such a requirement can obviously be relaxed considerably as far as practice is concerned. Pronounced skew and kurtosis in ability distributions are only occasionally damaging to the PROX calibration algorithm. As Table 6 shows, with 20 normal items there are no large Δ_i values for the distribution with mean of 3 and standard deviation of 1, despite the average raw score skewness and kurtosis of -2.00 and 8.50. Simulation shows no evidence that PROX would be misleading in any ordinary calibration situation, even when the assumption of normal distributions is clearly unmet.

Comparison of UCON and PROX for Accuracy--Real Data

Although we do not have any idea of the actual parameter values in real data, it will be informative to make a comparison between the UCON and PROX methods in these cases. We have selected eight calibration problems, seven of which were carried out independently by students in a course on

Educational Measurement at the University of Western Australia. Part of the course requirements was calibration of a test of the students own choosing. These seven examples vary considerably in the first four moments of their raw score distribution and hence represent a variety of typical situations.

The eighth example was a 40-item test constructed by Douglas for students registered in an introductory Measurement and Evaluation course at the same university.

Of the seven student examples, four produced UCON and PROX estimates which were identical for all practical purposes. One of the remaining three distributions was highly skewed; in this case, only one item in 30 was "poorly" estimated with the difference between UCON and PROX being 0.19. The other two distributions had moderate skew and in each case three items were "poorly" estimated. The maximum discrepancy observed was 0.35 but this was for an item on which the proportion of persons who had the item correct was 0.98.

Table 7 displays the proportion correct, the UCON and the PROX estimates ordered according to the magnitude of the UCON estimate for each item in the teacher-constructed test. Raw score statistics are reported for the distribution of scores. Despite the moderate negative skew and the fact that these are unedited items, not one item shows an absolute discrepancy between calibration methods of more than 0.07.

On the evidence of these eight real-data examples it

would appear that the major contributing factor to inaccurate estimation is not departure from normality but extreme item scores. Items with extreme scores are poorly estimated under even the theoretically ideal "conditional" method. These items have the largest standard errors of estimation. Calibration samples should in general be edited for both extreme items and extreme persons prior to submission to an estimation algorithm.

So far we have not mentioned the estimates of the standard errors in any of these examples. The PROX procedure is even more accurate for standard errors than for item difficulties. For example, in the student example mentioned above where a discrepancy of 0.35 was observed in item estimates, the discrepancy in standard errors was 0.17. When we compare this with the magnitude of the UCON estimate for this standard error of 1.02 a discrepancy of 0.17 seems negligible. The same conclusion applies in all other examples, both simulated and real-data.

Conclusions

We have introduced a new item calibration technique based on normal distribution assumptions in both the item arrangement and the person population. The technique is so simple it can be done by hand. The results from extensive simulations and from real examples indicate that one

may depart considerably from these assumptions before getting into a situation where PROX would be misleading as a calibrating technique. If reasonable editing of data is carried out before exposing the data to a calibration algorithm, it is difficult to envisage any calibration problems for which PROX will not be as efficient and as useful as the unconditional method.

Table 1

Logits from Proportions

Proportion	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
.0	-	-4.60	-3.89	-3.48	-3.18	-2.94	-2.75	-2.59	-2.44	-2.31
.1	-2.20	-2.09	-1.99	-1.90	-1.82	-1.73	-1.66	-1.58	-1.52	-1.45
.2	-1.39	-1.32	-1.26	-1.21	-1.15	-1.10	-1.04	- .99	- .94	- .90
.3	- .85	- .80	- .75	- .71	- .66	- .62	- .58	- .53	- .49	- .45
.4	- .40	- .36	- .32	- .28	- .24	- .20	- .16	- .12	- .08	- .04

logit $\ell = \ln [P/(1-P)]$

for proportions above .50 use 1-P and change the logit sign to +.

Table 2

An Example of the Hand Algorithm for PROX

p	l_i	① g_i	② $l_i g_i$	④ d_i^0	⑦	⑨ $d_i = Y d_i^0$
.1	2.20	1	2.20	1.83	Sum Top 6th=4.03	2.14
.2	1.39	2	2.78	1.02		1.19
.3	.85	4	3.40	.48		.56
.4	.41	5	2.05	.04		.05
			10.43			
.5	.00	4	.00	- .37	Sum Bottom 6th=-4.02	- .43
.6	- .41	2	- .82	- .78	S=8.05x2/19 = .85	- .91
.7	- .85	1	- .85	-1.22	V_d = .72	-1.43
.8	-1.39	1	-1.39	-1.76	D = .72/2.89 = .25	-2.06
.9	-2.20	0	.00	-		-
		20	-3.06			

Sum	7.37
③ Mean	.37

G = .28x.25 = .07

Y=(1.28/.93)$^{1/2}$=1.17

⑧ X=(1.25/.93)$^{1/2}$=1.16

f	⑤ b_f^0	n_f	⑥	⑩ $b_f = X b_f^0$
.1	-2.20	2	Sum Bottom 6th = -10.42	-2.55
.2	-1.39	5		-1.61
.3	- .85	6		- .99
.4	- .41	7		- .48
.5	.00	6	Sum Top 6th = 6.32	.00
.6	.41	6	S = 16.74x2/37 = .90	.48
.7	.85	4	V_b = .81	.99
.8	1.39	2	B = .81/2.89 = .28	1.61
.9	2.20	0		2.55

Table 3

Coefficient for Standard Errors of Calibration or Measurement

Standard Deviation of Ability in Sample or Difficulty in Test in Logits*	Proportion Correct on Item in Sample or Proportion Correct by Person on Test					
	.50 .50	.40 .60	.30 .70	.20 .80	.10 .90	.06 .94
0.25	2.0	2.1	2.2	2.5	3.3	4.2
0.50	2.1	2.1	2.3	2.6	3.4	4.2
0.75	2.2	2.2	2.3	2.6	3.5	4.3
1.00	2.3	2.3	2.4	2.7	3.6	4.4
1.25	2.4	2.5	2.6	2.8	3.7	4.5
1.50	2.6	2.6	2.7	2.9	3.7	4.5
1.75	2.7	2.7	2.8	3.1	3.8	4.5
2.00	2.9	2.9	3.0	3.2	3.9	4.6

The standard errors of calibration (or measurement) equal the tabled C coefficient divided by the square root of the sample size, $SE(d_i) = C_i/N^{1/2}$ or the square root of test length, $SE(b_f) = C_f/L^{1/2}$

* in the values of the estimation algorithm:

Person ability standard deviation = $1.7YD^{1/2}$

Item difficulty standard deviation= $1.7XB^{1/2}$

Table 4

Item Difficulty Estimates Based on UCON,
Computer PROX and Hand PROX Methods

	Estimates		
Item Number	UCON	Computer PROX	Hand PROX
1	2.09	2.20	2.14
2-3	1.22	1.24	1.19
4-7	0.60	0.60	0.56
8-12	0.07	0.06	0.05
13-16	-0.42	-0.43	-0.43
17-18	-0.93	-0.94	-0.91
19	-1.50	-1.51	-1.43
20	-2.23	-2.27	-2.06

Table 5

Averages over Four Replications of the PROX and UCON Estimates of the Uniform Items 1-5 and 16-20 Administered to $\beta \sim (2,2)$

Item Number	G_i	$\bar{P}_i$	$\bar{P}_i - G_i$	$\bar{U}_i$	$\bar{U}_i - G_i$	$\Delta_i = \lvert\bar{P}_i - \bar{U}_i\rvert$
1	-1.9	-1.75	.15	-1.86	.04	0.11
2	-1.7	-1.59	.11	-1.70	.00	0.11
3	-1.5	-1.47	.03	-1.57	-.07	0.10
4	-1.3	-1.34	-.04	-1.44	-.14	0.10
5	-1.1	-1.10	.00	-1.19	-.09	0.09
16	1.1	1.16	.06	1.25	.15	-0.09
17	1.3	1.24	-.06	1.34	.04	-0.10
18	1.5	1.33	-.17	1.45	-.05	-0.12
19	1.7	1.54	-.16	1.68	-.02	-0.14
20	1.9	1.89	-.01	2.04	.14	-0.15

Average Raw Score Skewness = -0.91

Average Raw Score Kurtosis = 2.92

G_i is the generating parameter

$\bar{P}_i$ is the average PROX estimate

$\bar{U}_i$ is the average UCON estimate

The trend in Δ_i is atypical.

Table 6

Summary of Simulation Results for Various D Values with Four Tests and Twelve Distributions

Number of Items	Item Distribution					
	Normal			Uniform		
	$.1 < \Delta_i \leq .2$		$.2 < \Delta_i \leq .3$	$.1 < \Delta_i \leq .2$		$.2 < \Delta_i \leq .3$
20	$\beta \sim (0,2)$	3*	-	$\beta \sim (2,\cdot 5)$	2	-
	$\beta \sim (1,2)$	1	-	$\beta \sim (2,1)$	1	-
	$\beta \sim (3,2)$	1	-	$\beta \sim (3,1)$	2	-
				$\beta \sim (0,2)$	2	-
				$\beta \sim (1,2)$	2	-
40	$\beta \sim (2,\cdot 5)$	1	-	$\beta \sim (2,\cdot 5)$	2	-
	$\beta \sim (2,1)$	1	-	$\beta \sim (2,1)$	2	1
	$\beta \sim (3,1)$	2	-	$\beta \sim (3,1)$	3	3
	$\beta \sim (0,2)$	1	-	$\beta \sim (0,2)$	2	-
				$\beta \sim (1,2)$	2	-

*The number of items with Δ values of given magnitude.

Table 7

A Comparison of UCON and PROX estimates for a Teacher-Constructed Test

Item Number	Proportion Correct	UCON	PROX	Item Number	Proportion Correct	UCON	PROX
3	.96	-2.32	-2.39	5	.68	.04	.05
22	.92	-1.58	-1.61	20	.68	.04	.05
7	.91	-1.51	-1.54	11	.67	.07	.07
25	.90	-1.45	-1.48	37	.66	.17	.18
33	.89	-1.33	-1.35	35	.63	.29	.30
23	.87	-1.12	-1.13	28	.61	.37	.37
8	.85	-0.98	-0.99	26	.59	.46	.47
34	.84	-0.85	-0.86	38	.58	.53	.54
12	.83	-0.77	-0.78	31	.57	.56	.56
29	.82	-0.73	-0.74	6	.57	.58	.59
2	.82	-0.73	-0.74	15	.54	.70	.70
27	.82	-0.73	-0.74	32	.53	.74	.75
18	.82	-0.73	-0.74	13	.52	.79	.79
40	.80	-0.62	-0.63	30	.51	.81	.82
4	.77	-0.42	-0.42	1	.48	.95	.96
17	.76	-0.36	-0.36	36	.41	1.22	1.23
16	.72	-0.15	-0.15	21	.35	1.51	1.53
14	.71	-0.10	-0.09	10	.33	1.62	1.63
9	.71	-0.10	-0.09	24	.20	2.29	2.32
19	.69	.01	.02	39	.13	2.81	2.86

Person Score: Mean = 26.65 Standard Deviation = 4.15 Skewness = -0.68
Kurtosis = 3.01

APPENDIX

The program steps in FORTRAN required for obtaining the PROX estimates of item parameters and standard errors are shown below.

```
C     N, L, S(I) AND R(J) HAVE BEEN READ
C     OBTAIN INITIAL CENTERED ITEM ESTIMATES, STORE IN VECTOR D
      CENTER = 0.0
      DO 2 I = 1,L
      D(I) = ALOG(N-S(I))/S(I))
    2 CENTER = CENTER + D(I)
      DO 3 I = 1,L
    3 D(I) = D(I) - (CENTER/L)
C     OBTAIN INITIAL ABILITY ESTIMATES, STORE IN VECTOR B
      L1 = L - 1
      DO  1 J = 1, L1
      ABILT = J
    1 B(J) = ALOG(ABILT/(L-ABILT))
C     FIND MEAN AND VARIANCE OF INITIAL ABILITIES AND THE
C     VARIANCE OF INITIAL DIFFICULTIES.
      AMEANB = 0.0
      VB = 0.0
      DO 4 J = 1,L1
      AMEANB = AMEANB + B(J)*R(J)
    4 VB = VB + B(J)*B(J)*R(J)
      VB = (VB-AMEANB*AMEANB/N)/(N-1)*2.89
      VD = 0.0
      DO 6 I = 1,L
```

```
  6  VD = VD + D(I)*D(I)
     VD = VD/L1*2.89
C    CALCULATE EXPANSION FACTORS
     G = VB*VD
     AB = SQRT [(1.0 + VB)/(1.0 - G)]
     AD = SQRT [(1.0 + VD)/(1.0 - G)]
C    FIND ITEM DIFFICULTIES AND THEIR ERRORS
     DO 7 I = 1,L
     D(I) = AB*D(I)
  7  SD(I) = SQRT(AB*N/ S(I)*(N-S(I))
C    FIND SCORE ABILITIES AND THEIR ERRORS
     DO 8 J = 1,L1
     B(J) = AD*B(J)
     ABILT = J
  8  SB(J) = SQRT (AD*L/ABILT*(L-ABILT))
```

REFERENCES

Andersen, E.B. The numerical solution of a set of conditional estimation equations. The Journal of the Royal Statistical Society: Series B, 1972, 34 (1), 42-54.

Cohen, L. A modified logistic response model for item analysis. (Unpublished manuscript), 1976.

Lord, F.M. A survey of observed test-score distributions with respect to skewness and kurtosis. Educational and Psychological Measurement, 1955, 15, 383-389.

Lord, F.M. An analysis of the verbal scholastic aptitude test using Birnbaum's three-parameter logistic model. Educational and Psychological Measurement, 1968, 28, 989-1020.

Lord, F.M. and Novick, M.R. Statistical Theories of Mental Test Scores. Reading, Massachusetts: Addison-Wesley, 1968.

Mason, G.P. and Odeh, R.E. A short-cut formula for standard deviation. Journal of Educational Measurement, 1968, 5 (4), 319-320.

Rasch, G. Probabilistic Models for Some Intelligence and Attainment Tests. Copenhagen, Denmark: Danmarks Paedagogiske Institut, 1960.

Rasch, G. An item analysis which takes individual differences into account. British Journal of Mathematical and Statistical Psychology, 1966, 19, 49-57.

Sabers, D.L. and Klausmeier, R.D. Accuracy of short-cut estimates for standard deviation. Journal of Educational Measurement, 1971, 8 (4), 335-339.

Wright, B.D. Sample-free test calibration and person measurement. in Proceedings of the 1967 Invitational Conference on Testing Problems. Princeton: Educational Testing Source, 1968.

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

Wright, B.D. and Douglas, G.A. Best test design and self-tailored testing. Research Memorandum, No. 19. Statistical Laboratory, Department of Education, University of Chicago, 1975(a).

Wright, B.D. and Douglas, G.A. Better procedures for sample-free item analysis. Research Memorandum, No. 20. Statistical Laboratory, Department of Education, University of Chicago, 1975(b).