Abstract

It is important to unambiguously identify the perceived level of difficulty of individual questions employed within a multiple-choice examination. The aim is to assign questions to one of three levels of difficulty. The number of question allocated to each level is an essential element in mapping the raw mark onto the standard University scale.

1. Introduction

Rasch analysis is a widely employed statistical technique. Its use has been reviewed [1] and covered in texts [2]. Interest here centres on the ability of the model to quantify the perceived level of difficulty of a series of multiple-choice questions. This identifies groups to which the questions belong and forms a key element in the scaling procedure currently adopted in this department.

The data are more fully described in the next section.

2. The Data

The examination considered here was a multiple-choice paper consisting of 100 questions sat by some 154 Stage I Psychology students at Newcastle University in the academic year 2008/9. The raw marks had mean 63.62, standard deviation 12.92 and resulted in frequencies (7, 14, 32, 45, 56) in the degree categories from fail to first class. The level of difficulty of each question was identified, and nominally labelled 40, 55 or 70. These are indicative of degree grades (third, lower second or first) rather than precise numerical values. It is this segregation of the questions that will be identified below. The lecturer concerned allocated questions to each category, it is claimed there were (34, 33, 33) in each band respectively. These values are employed in scaling based on the work of Angoff [3], which was used to map the raw mark totals onto the standard University scale. The resulting values were mean 58.68, standard deviation 8.87 and resulted in frequencies (1, 20, 62, 58, 13) in the degree categories from fail to first class. It should be noted that this approach only depends on the number of questions in each category (40, 55, 70), not the precise category a question belongs to.

The model is now introduced.

3. The Model

The key assumptions and model have been previously described [1]. It is assumed that for each person \(i: i=1,\ldots,N\) taking the examination there is a parameter, \(\theta_i\), measuring
their ability, and for each question (item) \( j: j = 1, \ldots, n \) there is a parameter measuring its difficulty \( b_j \). Where, for convenience, the notation follows that of Hirose [4].

Given a specific person and question, the Rasch model gives the probability that person \( i \) answers item \( j \) correctly by \( P_{ij} \) where

\[
P_{ij} = \frac{e^{\theta_i - b_j}}{1 + e^{\theta_i - b_j}}
\]

In a recent paper Hirose [4] investigated the item response theory, employing a very similar model. This was employed to assess the abilities of the examinees along with the difficulties of the problems (questions) it was

\[
Q_{ij} = \frac{1}{1 + e^{-1.75(\theta_i - b_j)}} = \frac{e^{1.75(\theta_i - b_j)}}{1 + e^{1.75(\theta_i - b_j)}}
\]

where \( Q_{ij} \) is the probability that student \( i \) gives the correct answer to problem \( j \). Here \( \theta_i \) represents the ability of student \( i \), \( a_i \) is a discrimination parameter and \( b_j \) a difficulty parameter associated with question \( j \). The student's ability, as reported by the model, is assumed to be positively correlated to the student's true aptitude. A more discriminating question is associated with a larger value of \( a_i \), while increasing levels of difficulty are associated with larger values of \( b_j \).

The likelihood for all students \( i = 1, 2, \ldots, N \) and all problems \( j = 1, 2, \ldots, n \) is

\[
L = \prod_{i=1}^{N} \prod_{j=1}^{n} P_{ij}^{Q_{ij}} (1 - P_{ij})^{1 - Q_{ij}}
\]

where \( Q_{ij} \) is 1 if the student answers the problem correctly and 0 otherwise. In the example considered here, and previously described, \( N = 154 \) and \( n = 254 \). By maximising \( L \) the maximum likelihood estimates of the parameters may be obtained. This amounts to estimating some \( n + N \) values, in this case some 254 values. With a basic programme and an antiquated computer the maximisation took only a few minutes to complete. As a cross check the DOS software Bigsteps (2.82) was employed, a Windows version Winsteps (3.71.0) [5], may be purchased. The solutions obtained were identical, up to a scale factor with those obtained here.

Having identified the levels of difficulty \( (b_j) \), can the questions be grouped?

### 4. Identifying Groups For The Level Of Difficulty

The first step is simply to examine the raw data. A probability plot (Fig 1) is informative.

![Fig 1](image-url)

**Fig 1 – Probability Plot Of The Level Of Difficulty**

So the test supports the hypothesis of normality when the p-value is greater than \( \alpha \), which is the case here.

Since the assumption of normality appears plausible, the next step is to see if the data can be explained by the sum of three normal distributions (for the assumed levels of difficulty). It is expected that the difficulty \( (b) \) follows a normal distribution \( b \sim a_i \phi (\mu_i, \sigma_i^2) + a_i \phi (\mu_j, \sigma_j^2) + a_i \phi (\mu_k, \sigma_k^2) \), with the condition that \( 1 = a_i + a_j + a_k \), \( 0 \leq a_i \leq 1 \) for the three weights. In addition \( \mu_i < \mu_j < \mu_k \), with strict inequality enforced.

The problem of estimating the parameters, which determine a mixture of densities, has been the subject of a large, diverse body of literature spanning nearly ninety years [6]. The general problem is formulated for a mixture of \( m \) normal distributions with parameters \( \mu_i \) and \( \sigma_i \) with density \( (pdf) \)

\[
\text{pdf} (x) = \sum_{i=1}^{m} a_i \phi (x, \mu_i, \sigma_i^2)
\]

where each \( a_i \) is non-negative and \( \sum_{i=1}^{m} a_i = 1 \).

Of interest here is the log-likelihood function [6], which can be maximised to estimate the parameters.

\[
L_i = \sum_{i=1}^{n} \log (\text{pdf} (x_i))
\]

As an alternative a least squares approach is adopted. The first step is to rank the observations so that \( x_{(i)} \leq x_{(j)} \leq \ldots \leq x_{(n)} \). Then minimise

\[
L_0 = \sum_{i=1}^{n} \left( \frac{i}{n} - \text{cdf} (x_{(i)}) \right)^2
\]

Where the cumulative probability density \( (cdf) \) is

\[
\text{cdf} (x) = \sum_{i=1}^{m} a_i \phi (x, \mu_i, \sigma_i^2)
\]

This approach is employed to provide a more ‘visual’ alternative.

The resulting parameters are presented in Table 1 overleaf. While the estimates differ, the allocation of question difficulty to the three distributions are \( (52, 1, 47) \) for \( L_0 \) and \( (51, 2, 47) \) for \( L_i \), indicating that the questions do not fall into three levels of difficulty. This shortcoming has
implications for the scaling employed [3]. The following graphs (Figs 2 and 3) indicate why the distribution only, effectively, provides a binary division.

Here $f_1$, $f_2$ and $f_3$ represent the three, appropriately weighted, normal distributions. When using the cumulative plot, a difficulty ($b$) is located on the abscissa, and then its level of difficulty group may be identified by projecting up to the cumulative curve. Only the extreme upper tail values are allocated to the second group.

An investigation of the levels of difficulty allocated by the lecturer was also conducted. As the preceding investigating suggests, this was uninformative. To briefly summarise the results, while those questions at level ‘40’ were correctly identified as a lower level of difficulty, those at ‘55’ and ‘70’ were effectively indistinguishable.

### 5. Conclusions

The method described provides an algorithmic approach for the identification of the level of difficulty of multiple-choice examination questions. When the results are scaled to fit the standard University scale the number of questions assigned to each level of difficulty is a key parameter [3]. It is clear that in this case the assumed separation with approximately equal numbers of questions in each group is fallacious. It is suggested that the calculated question difficulties should be fed back to the lecturer concerned, to assist in meeting the desired split into three groups, or more probably an alternate scaling approach should be adopted.

### References


### About The Author

Mike Cox is a lecturer specialising in numerate disciplines.