

## Chapter 3

# Statistical Analyses of Data from the IEA Reading Literacy Study

*Peter Allerup*

The Danish National Institute for Educational Research, Copenhagen

### 1. Introduction

The Reading Literacy Study took place in the years 1990-91 as an international project under IEA, the International Association for the Evaluation of Educational Achievement. Important objectives for the project were to develop three international reading scales for assessing reading abilities and to study systematic background factors which could influence high or low reading abilities of students in grades 3/4 and grade 8. Many national studies that aim at finding reasons for high or low reading achievement fail to come up with clear explanations, as the variation of background factors within a country is very limited, a result of homogeneous educational systems. The background variability has obvious consequences for the study of achievement as a dependent variable in relation to independent background variables, but this variability also provides a powerful basis for the analysis of homogeneity across countries of the reading scales.

The study was conducted as a pilot in 1990 and a main study in the years 1990-91 on two different populations of students: 9-year-olds from grades 3/4 (pop A) and 14-year-olds (pop B) from grades 8/9 in thirty-two countries around the world.

The study is a continuation of earlier IEA reading studies (Thorndike, 1973); it was not based on a priori formulated reading hypotheses. The data were collected in all thirty-two countries according to already accepted sampling procedures and were then analyzed in an International Coordinating Center. Conclusions drawn from data could be considered as representative for students/teachers in the countries at the specified grades.

The test materials were comprised of test booklets to the students, questionnaires to students, questionnaires to teachers and school principals and, on local initiative, a number of „national options“ added to test booklets and questionnaires. The first international statistical analyses based on all data have been published: „How in the world do students read?“ (Elley, 1992), „Teaching reading around the world“ (Lundberg and Linnakyla, 1992) and „Effective Schools in reading“ (Postlethwaite and Ross, 1992).

The student test booklets consisted of text passages which were presented to the students and after having read a passage, a number of questions - which are from now on the items of the test - were then presented in either closed multiple response-category form (mc) or as open ended questions. The mc-questions were constructed, usually, with four answer categories, and one correct answer.

The text passages were a priori grouped into three domains. Items within one domain were all expected to draw upon one latent reading ability:

- *Narrative* text type: Texts that tell a story or give the order in which things happen (pop A: 22 items, pop B: 29 items); relating to the most common reading ability.
- *Expository* text types: Texts that describe things or people or explain how things work or why things happened (pop A: 21 items, pop B: 26 items); relating to reading ability, where e.g., conclusions are drawn from the text.
- *Document* text types: Tables, charts, diagrams, maps (pop A: 23 items, pop B: 34 items); relating to reading abilities where the input is not a usual sequence of written words.

From a measurement point of view, the intention behind the construction of three domains was to calculate and compare student abilities by means of single-valued indices for each domain. The technical and methodological steps taken during the statistical analyses which finally lead to the construction of these three International Reading Scales and the problems of squeezing the multidimensional item information to three one dimensional indices will be summarized.

Much effort was devoted to the task of making the test booklets for each population A and B to work as one common test across all countries. Besides problems of proper translations into thirty-two languages this phase of the test construction involved problems like layout, illustrations to the text, instructions to students and teachers and definition of time limits for the tests.

A fundamental aim of the statistical analyses was to select those items from the original pool that were equally difficult for all students around the world, and include these in the three International Reading Scales.

On the one hand this seems legitimate, since this would, indeed lead to a common ruler for all students when measuring student abilities. On the other hand, one might argue, whatever the item difficulties, in fact, are, it should be possible by appropriate statistical modelling to utilize data on all items and then, based on a relevant statistical model, to calculate a measure of ability for each student. According to this view the construction of the International Reading Scales is first of all a matter of modelling the entire data set of item responses, exactly as they are and, then, directed by the statistical model, estimate student ability.

By means of percentage correct responses, calculated item by item, and methods based on correlations, classical test theory evaluate the appropriateness of each item as member of a reading scale. Such criteria for item selection depend on the thirty-two samples and do not directly address the issue that the number of correct responses for each domain, the student scores, can be used to measure student ability.

The logistic IRT (Item Response Theory) models, and especially the one parameter logistic Rasch model include item difficulties and student abilities as systematic parts of the statistical model, and use the total number of correct responses as statistics for item difficulties and student abilities. This is, in fact, tested under the Rasch model by testing statistical sufficiency of these statistics. It is a virtue of the Rasch model that any investigation concerning item difficulties can be undertaken independent of student abilities. This is important when a set of item difficulties are evaluated across the thirty-two populations.

The statistical analyses pointed to items that were dropped from all further analyses, to items that were accepted for within-country comparisons, and to items that could be used across all countries.

## 2. IRT-models

The data from the two IEA populations were analyzed separately. It was accepted as a consequence of the very construction of items for the three domains: Narrative, Expository and Document that the responses from the three domains were analyzed separately. The three student abilities estimated on this background were, in the international reports pooled to on average value, a kind of total reading ability.

The data  $((x_{vi}))$  from each domain is a  $n$  by  $k$  matrix of observations with  $x_{vi} = 1/0$  if the response is correct/noncorrect. Such a matrix exists for each of the participating thirty-two countries.

The one-parameter model (Rasch, 1960), which is the simplest of the one-, two- and three parameter IRT logistic models, assumes a set of item parameters  $\sigma_1, \dots, \sigma_k$ , measuring the difficulty of the items (or easiness depending on the sign of the parameters), and a set of student parameters  $\theta_1, \dots, \theta_N$ , measuring the abilities of the students.

Given these two sets of parameters, the Rasch model represents the probability of a correct response ( $X_{vi} = 1$ ) to item  $i$  with difficulty  $\sigma_i$  for a student with ability  $\theta_v$  by the following equation (see chapter 1.1):

$$p(X_{vi} = 1) = \frac{e^{\theta_v - \sigma_i}}{1 + e^{\theta_v - \sigma_i}}. \quad (1)$$

The model (1) is a particularly powerful member of the class of IRT models, because it can be shown (Rasch, 1968) that the mathematical structure unambiguously reflects two properties (a): the raw scores for the items (i.e., the number of correct answers across all items) and the raw scores for the students (i.e., the total number of correct answers per student) are jointly sufficient statistics for the item and student parameters and (b): it is possible to compare student abilities irrespective of which items (difficulties) have been used for the comparisons. The first property is a statistical „validation“ of the practical use of raw scores; the second, which is due to Rasch, called specific objectivity, has a symmetric requirement for comparisons between item difficulties irrespective of student abilities.

The two properties are connected, because statistical sufficiency under the Rasch model means that inference concerning difficulties of the items can be investigated independent of the student abilities  $\theta_1, \dots, \theta_N$ , if conditional distributions of responses given the student totals  $r_v$  are analyzed.

Seen from a technical point of view, the Rasch model (1) is a model in which all Item Characteristic Curves (ICC) have the same slopes and are equally discriminating between the students.

$$ICC = p(X_{vi} = 1 | \sigma_i, \theta) \quad (2)$$

It has obvious statistical advantages if the one-parameter model (1) can be fitted to data, but frequently the data do not satisfy these requirements because varying slopes of the

estimated ICC's are observed in real data. One way of dealing with this problem is by adopting a generalization of the one-parameter Rasch model, the two-parameter logistic model (Lord and Novick, 1968), (see chapter 1.3):

$$p(X_{vi} = 1) = \frac{e^{\beta_i(\theta_v - \sigma_i)}}{1 + e^{\beta_i(\theta_v - \sigma_i)}} \quad (3)$$

In model (3) the item discrimination parameters  $\beta_1, \dots, \beta_k$  are the slopes of the ICC's and they must be estimated from data together with the  $\theta$  - and  $\sigma$  parameters. It is seen that the one-parameter model (1) is formally obtained from the two-parameter model (3) by setting  $\beta_i = 1$  for all  $i$ .

Although the two-parameter model has been widely used (Lord, 1980) the model (3) suffers from both theoretical and practical problems and limitations. Although the mathematical structure of the Rasch model is a special case of the two parameter model, when  $\beta_i = 1$ , all  $i$ , it can be discussed, what is, in fact generalized from model (1) to model (3) in terms of interpretation.

First, it is not possible under the two parameter model to give independent interpretations of the item and discrimination parameters; this can be seen from attempts to identify the components  $\beta_i$ ,  $\theta_v$  and  $\sigma_i$ , which control the probability completely. The combined additive term  $\theta_v - \sigma_i$  interacts multiplicatively with the  $\beta_i$ . Hence, by „shrinking“ the combined scale of difficulty and ability  $\theta_v - \sigma_i$ , i.e., dividing by some constant, this can be „compensated“ by multiplying the  $\beta_i$ 's by the same constant, and the final value of the probability,  $p(x_{vi})$ , remains unchanged.

Second, the discrimination parameters  $\beta_1, \dots, \beta_k$  are usually unknown in empirical data analyses. This brings the model (3) outside the class of exponential distributions, which among other things implies severe estimation problems of the parameters, including the  $\beta$ 's.

If one proceeds with test items under model (3) having varying  $\beta_i$  values, some of which being different from unity, it should be kept in mind that the estimate of student ability  $\theta_v$  is based on the statistic  $T_v = \sum \beta_i x_{vi}$ . This statistic represents the sum of item discriminations across those items which are answered correctly ( $x_{vi} = 1$ ). Increasing  $T_v$  values, and thereby higher ability  $\theta_v$  estimates, are, therefore, a consequence of either responding correctly to items with high  $\beta$ 's, possibly being easy items with low  $\sigma$ 's, or responding correctly to an increasing number of items.

This illustrates the confusion in the two parameter model concerning the interpretation of the ability parameter  $\theta_v$  where the Rasch model unambiguously combines the student ability with the number of correct answers, i.e., the total score  $r_v$ .

Sometimes it can be argued that a response system with closed response categories may give rise to guessing, especially in case of poorly performing students.

During the statistical analyses of the IEA data it was, therefore, briefly discussed to introduce the three-parameter model, which makes use of a parameter  $\phi$ , measuring the probability of obtaining a correct response by chance. By mixing this probability,  $\phi$ , with the usual two-parameter model, or alternatively a one-parameter model, one gets the three parameter model:

$$p(X_{vi} = 1) = \frac{e^{\beta_i(\theta_v - \sigma_i)}}{1 + e^{\beta_i(\theta_v - \sigma_i)}}(1 - \varphi) + \varphi. \quad (4)$$

It is seen that  $\varphi = 0$  brings the three parameter model (4) back to the two parameter model (3), like  $\beta_i = 1$ , all  $i$ , reduces the two parameter model to the Rasch model.

The three parameter model is, like the two parameter model, outside the family of exponential distributions, which causes serious mathematical statistical problems for the estimation of parameters and testing the fit. Furthermore, it can be shown, that this model exhibits even stronger interpretation problems than the two parameter model regarding the structure of model and explanation of parameters.

### 3. Analytic strategy - data analysis

At a first glance it seems like three theoretical frameworks in terms of three statistical models are available for the statistical analysis of the test book data, viz. the one-, two- and three parameter logistic models. Furthermore, one could claim that the models are ordered hierarchically taking the generalization from the one parameter model to the three parameter model outlined above as the ordering, i.e., including more and more aspects of the response process. However, the interpretation of e.g., item difficulties changes from the one parameter model to the two- and three parameter model. Consequently, the apparently more elastic two- and three parameter models show growing difficulties in other respects compared to the one parameter model.

From the initial data analyses it became clear that the item characteristic curves, ICC, had varying slope values, some of them significantly different from unity. Furthermore, it was surprising to witness a pattern, when an item revealed a slope value for the ICC distinctly different from unity, this value remained constant across all data sets from different, independent countries. The impression was, therefore, that the observed deviation from unity could not simply be ascribed to chance.

These empirical observations do, in fact, indicate that the construction of an International Reading Scale would be difficult, if only the criteria behind the one-parameter model are used. Thus, it was discussed, whether priority should be given to attempts on fitting the one-parameter model, at the cost of losing a number of misfitting items, or to go on with two-parameter analysis and accept the described draw backs of this model.

A method is, however, available through the characterization of the one-parameter model (Allerup, 1994). In fact, the correspondence between the Rasch model and the sufficiency of total scores  $n_i$  (for items) and  $r_v$  (for students) makes it possible, through conditional inference, within the Rasch model, to investigate features like unequal ICC-slopes as alternatives to the model under hypothesis. In other words, using the one-parameter model as a platform, test statistics under the Rasch model can: (a) point to ICC's with slopes different from unity and, (b) estimate this slope by means of estimators related to the simple Rasch model. This procedure is justified by the following analytic steps.

The probability of a correct response to item  $i$  conditional on the student score  $r$ ,

$$\begin{aligned}
p(X_{vi} = 1|r_v) &= \varepsilon_i \frac{\gamma_{r-1}^{(i)}}{\gamma_r^{(i)}} = \frac{e^{\sigma_i + \log(\gamma_{r-1}^{(i)}/\gamma_r^{(i)})}}{1 + e^{\sigma_i + \log(\gamma_{r-1}^{(i)}/\gamma_r^{(i)})}} \\
\sigma_i &= \log(\varepsilon_i) \quad \gamma_r = \sum_{\omega} \prod \varepsilon_i^{x_i} \\
\omega &= (x_1, \dots, x_k | \sum x_i = r)
\end{aligned} \tag{5}$$

can be identified with the unconditional probability of the Rasch model (1).

Thus, the two parameter generalization of the one-parameter model can be captured directly from (5) by the following conditional model structure

$$\begin{aligned}
p(X_{vi} = 1|r_v) &= \frac{e^{\sigma_i + \beta_i \lambda_{ri}}}{1 + e^{\sigma_i + \beta_i \lambda_{ri}}} \\
\lambda_{ri} &= \log(\gamma_{r-1}^{(i)}/\gamma_r^{(i)})
\end{aligned} \tag{6}$$

where  $\beta_i$  plays the role of item discrimination parameter and the  $\gamma$ 's are the elementary symmetric functions of the item difficulties  $\varepsilon_1, \dots, \varepsilon_k$  (Rasch, 1960). The  $\gamma^{(i)}$ 's are the same as  $\gamma$ , but the function is calculated without item difficulty  $\varepsilon_i$ .

The reason for suggesting  $\sigma_i + \beta_i \lambda_{ri}$  rather than  $\beta_i(\sigma_i + \lambda_{ri})$  which would otherwise make the equivalence with the two parameter model (2) clear, is that (6) immediately represents a simple one-dimensional logistic regression of  $x_{vi}$  on the  $\lambda$ 's for fixed item  $i$  ( $r = 1, \dots, k - 1$ ) with slope =  $\beta_i$  and intercept =  $\sigma_i$ . Considering  $\lambda_{ri} = \lambda_{ri}(\sigma_1, \dots, \sigma_k)$  as known by the estimated  $\sigma$ 's, these quantities are by-products of solving the conditional maximum likelihood equations for  $\sigma_1, \dots, \sigma_k$  (Andersen, 1973) and tentative tests of fit for the logistic regression can be conducted. Beside the likelihood ratio test ( $\chi^2$ ) for linearity, the hypothesis  $H: \beta_i = 1$  of slopes equal unity is of interest, because  $\beta_i = 1$  (all  $i$ ) is, in fact, a test for the basic Rasch model. The number of degrees of freedom under  $H$  and the ( $\chi^2$ ) test is fixed by the number of different  $\lambda_{ri}$  values, and standard errors etc. are calculated using estimates obtained under the logistic regression.

This definition and analysis of item discrimination  $\beta_i$  takes place inside the model (6), which, like the model (5) derived from the Rasch model, is conditional on the statistics  $r_v$  for the student parameters  $\theta_1, \dots, \theta_N$  under the original Rasch model. By this we avoid the mathematical confusion caused by conditioning by the statistic  $T_v = \sum \beta_i x_{vi}$ , which is the sufficient statistic for the student parameter  $\theta_v$ , under the two parameter model, if the  $\beta$ 's are known.

The conclusion concerning the choice of a statistical model was that test statistics for the conditional two-parameter version (6) should be identified. Tentative likelihood ratio tests of linearity, i.e., the fit for the 'shape' of ICC, should be calculated, considering  $\lambda_{ri} = \lambda_{ri}(\sigma_1, \dots, \sigma_k)$  as known, and a test for the hypothesis  $H: \beta_i = 1$  should be undertaken. In the case that  $H$  was rejected, a test for constant item discrimination values across different countries should be carried out. By this procedure an item was allowed to discriminate by a value different from unity as long as it attained this value across all countries.

It was decided to measure the degree of guessing by comparing the observed and expected number of correct responses in the 25% lowest score group, calculated under the one- or two-parameter model. If the ratio observed/expected is high, it is an indication of guessing.

### *Test procedures and test statistics*

The statistical analyses of the data were administered according to the following program. Data from a given domain and a given population from a given country were first submitted to test of fit of the conditional two-parameter model.

As a result of screening each item by four National Test Statistics, every item was flagged according to: „yes“ = the test statistic accepts, „no“ = the test statistic rejects. The National Test Statistics are constructed to test for particular deviations from the conditional two-parameter model within a given country.

All items were then, one by one, viewed through International Statistics, and the items were flagged accordingly. The International Statistics evaluate structures across countries. By this procedure an item could be flagged by up to six times.

In a few cases it was found that a country, considered as one „unit“ showed unacceptably many flags across all items. In such cases it was decided to omit data from the country. In a few other cases it was found that a single item showed unacceptably many flags across all countries. In such cases, these items were left out from all subsequent analyses. Quite often it was found that the flagging procedure tended to concentrate for specific combinations of items and countries. In such events the items were dropped specifically for those countries.

### *NATIONAL STATISTICS*

1. DIST - **d**istance measure. For each item and for given score level the observed number of students with correct responses is compared with the expected number of students.
2. DIFINT - **d**ifference **i**nternal between item difficulties estimated from the low-scoring students and the high scoring students.
3. TESHAP - **t**est of **s**hape of the item characteristic curve. It is tested, using logistic regression, if the observed frequency of students responding correctly to an item is compatible with the one-parameter Rasch model or the conditional two-parameter model.
4. TEEXT - **t**est of **e**xternal variable, viz. test of sex bias. A chi-square statistic is calculated across the score groups from the series of 2 by 2 tables with score level and sex as entries.

### *INTERNATIONAL STATISTICS*

5. THETOT -**t**heta parameters based on the **t**otal data for each country, i.e., the item difficulties for the items in the particular domain under investigation. The estimated item difficulties  $\sigma_i$  are compared across all countries. This may lead to either dropping the item completely from all analyses or to the dropping of the item for specific countries.
6. ITDISCR -**i**tem **d**iscrimination. It is tested if  $\beta_i = 1$  for each item within each country cannot be rejected and, if rejected, it is further tested, if the  $\beta_i$  values are consistent across countries.

#### 4. Estimation of student abilities

The student abilities were estimated using the final pool of items accepted for the International Reading Scale. Having estimated the item difficulty and discrimination parameters, the maximum likelihood equations for the student abilities  $\theta_v$  are the following, considering the item parameters to be known by their estimated values:

$$r = \sum_{i=1}^k \frac{e^{\theta_v - \sigma_i}}{1 + e^{\theta_v - \sigma_i}} \quad r = 1, \dots, k - 1 \quad (7)$$

$$r = \sum_{i=1}^k \frac{e^{\beta_i(\theta_v - \sigma_i)}}{1 + e^{\beta_i(\theta_v - \sigma_i)}} \quad r = \beta_1, \dots, \sum \beta_j - \min(\beta_i). \quad (8)$$

For both the one- and the two parameter model *no* estimate of the ability ( $\theta_v$ ) exists for students with zero-score or full-score (i.e., all items correct); various 'randomization' procedures assigning  $\theta$ -values to such students were considered.

The evaluation of fit of a student response pattern was obtained using exact probabilities of the actual responses ( $x_{v1}, \dots, x_{vk}$ ) to the items *given* the student score  $r_v$ :

$$p_{\text{krit}} = \sum_{\omega} p((x_1, \dots, x_k) | r_v) \quad (9)$$

$$\omega = \left\{ (x_1, \dots, x_k) \mid p(x_1, \dots, x_k | r_v) \leq p(x_{v1}, \dots, x_{vk} | r_v) \right\}$$

If this probability falls below e.g., 0.05, one might consider the student response pattern aberrant.

#### 5. Results and conclusions

It turned out that only a few items could be sustained for all countries. Consequently, the set of items in the International Reading Scale must be considered a set of reference items instead of one fixed scale with fixed marks for measuring. The reason why this reference idea works successfully is that under the one parameter Rasch model any subset of a set of fitting items can be used for the estimation of student abilities. This property is approximately valid for the two parameter model also, provided that the item discriminations  $\beta_i$  are not very different from unity. The test statistics used in the flagging procedure were the following.

If one would have hoped for the finding of a few, important background factors which could account for most of the variance found in the student achievement levels around the world it turned out different (Elley, 1992). No single variable or variables can 'explain' why the students performed well or not so well; future analyses will, however continue investigating more details in the national data sets and may reveal structures which provide useful information to national educational planners.

## References

- Allerup, P. (1994). Rasch measurement, theory of. *The international encyclopedia of education*; second edition, Pergamon Press.
- Andersen E.B. (1973). *Conditional inference and models for measuring*. Copenhagen: Mentalhygiejnisk Forlag.
- Elley, W.B. (1992). *How in the world do students read*. The Hague: The International Association for the Evaluation of Educational Achievement.
- Lord F.M. (1980). *Applications of item response theory to practical testing problems*. New Jersey: Lawrence Erlbaum Associates.
- Lord F.M. and Novick M.R. (1968). *Statistical theories of mental test scores*. Massachusetts: Addison Wesley.
- Lundberg I. and Linnakyla P. (1992). *Teaching reading around the world*. The Hague: The International Association for the Evaluation of Educational Achievement.
- Postlethwaite T.N. and Ross K. (1992). *Effective schools in reading, implications for educational planners*. The Hague: The International Association for the Evaluation of Educational Achievement.
- Rasch G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research.
- Rasch G. (1968). *A mathematical theory of objectivity and its consequences for model construction*. Amsterdam: Paper read at European Meeting on Statistics, Econometrics and Management Science.
- Rasch G.(1971). *Proof that the necessary condition for the validity of the multiplicative dichotomic model is also sufficient*. Copenhagen: Dupl. note, Statistical Institute (see Allerup,1994).
- Thorndike, R.(1973). *Reading comprehension education in fifteen countries*. Uppsala: Almqvist & Wiksell.