AUTHOR Blais, Jean-Guy
TITLE Item Response Theory Scaling with Heterogeneous Populations.
PUB DATE Apr 93
NOTE 49p.; Paper presented at the Annual Meeting of the American Educational Research Association (Atlanta, GA, April 12-16, 1993).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Academic Achievement; Adolescents; Cross Cultural Studies; Data Analysis; Elementary School Students; Equated Scores; Estimation (Mathematics); Foreign Countries; International Studies; *Item Response Theory; Junior High Schools; *Junior High School Students; *Mathematics Tests; Preadolescents; Primary Education; *Reference Groups; Sampling; *Scaling; *Science Tests; Scores
IDENTIFIERS Diversity (Student); Item Parameters; Linkage; *Second International Assessment of Ed Progress; Three Parameter Model

ABSTRACT

Tools used in scaling proficiency scores from the Second International Assessment of Educational Progress (IAEP) are described. The second IAEP study, conducted in 1991, was an international comparative study of the mathematics and science skills of samples of 9- and 13-year-old students from 20 countries. This paper focuses on part of the second stage of data analysis, work done in creating a unique scale for all the participating populations, that is, creating reference populations, scaling methodology, and linkage of 9- and 13-year-old populations. All populations contributed to four combined reference populations, superpopulations, one for each age group by mathematics and science combination. The item response theory scaling model is the three-parameter logistic model. Linking was accomplished through a small set of common items included in the 9- and 13-year-old assessments in each subject area. Results show that even if on an item-by-item basis the equating gives results that are not ideal, when all items are taken into account, student proficiency scores estimated with both sets of item parameter estimates can be considered to be on the same scale. Results further indicate that no information has been lost as a result of using one set of item parameter estimates over another. Fourteen tables and 16 figures present analysis data. (SLD)

Item response theory scaling with heterogeneous populations

Jean-Guy Blais

University of Montreal

Paper presented at the annual meeting of the American Educational Research Association, Atlanta, april 1993.

Introduction

This paper describes the tools used in scaling proficiency scores from the second International Assessment of Educational Progress (IAEP)[1]. The reader already acquainted with the IAEP technical report volume two will find many similarities bewteen this paper and the technical report. It's normal, both are authored by the the same person and describe the same work. Moreover, the reader acquainted with NAEP technical report 17-R-20 will find that the same technical tools were used in NAEP and IAEP. The important difference is that these tools were used with much more heterogeneous populations than in any study done before: up to 36 different populations assessed in 13 different languages. This heterogeneity rendered the attempt to put proficiency scores on the same scale a delicate task wich must be undertaken carefully.

Since this paper inspired itself largely from the IAEP technical report volume two and that this report was written in collaboration, I would like to thank E.G. Johnson, R.J. Mislevy, P.J. Pashley, K.M. Sheehan and R.J. Zwick, all from Educational Testing Service, whose collaboration and experience were very precious. I would also like to thank Nancy Mead, Archie Lapointe and Jan Askew for their description of the IAEP populations that comes in the next section.

The Second International Assessment of Educational Progress

The second IAEP study, conducted in 1991, was an international comparative study of the mathematics and science skills of samples of 9- and 13-year-olds students from 20 countries. IAEP was designed to collect and report data on what students know and can do, on the educational and cultural factors associated with achievement, and on students' attitudes, backgrounds, and classroom experiences.

[1] The second International Assessment of Educational Progress was supported financially by the National Science Foundation and the U.S. Department of Education's National Center for Education Statistics for the expenses of overall coordination, sampling, data analysis, and reporting. The Carnegie Corporation provided additional funds to cover the travel expenses of some of the participants who could not meet the financial burdens of traveling to the project's coordination and training meetings held in Canada, England, France, Hong Kong, and the United States. Decisions concerning the design and implementation of the project were made in collaboration with the representatives of the countries involved in the survey. The National Academy of Sciences' Board on International Comparative Studies in Education reviewed plans for IAEP at several stages of its development and made suggestions to improve the technical quality of the study. The board is responsible for reviewing the soundness of the technical procedures of international studies funded by federal agencies of the U.S. government.

This project was a four part survey: a main assessment of 13-year-olds' performance in mathematics and science; an assessment of 9-year-olds' performance in mathematics and science; an experimental, performance-based assessment of 13-year-olds' ability to use equipment and materials to solve mathematics and science problems; and a short probe of the geography skills and knowledge of 13-year-olds. All countries participated in the main assessment of 13-year-olds; participation in the other assessment components was optional.

Some countries drew samples from virtually all children in the appropriate age group; others confined their assessments to specific geographic areas, language groups or grade levels. The definition of populations often followed the structure of school systems, political divisions, and cultural distinctions. For example, the sample in Israel focused on students in Hebrew-speaking schools, wich share a common curriculum, language and tradition. All countries limited their assessment to students that were in school, wich for some participants meant excluding significant numbers of age-eligible children. In a few cases, a sizable proportion of the selected schools or students did not participate in the assessment, and therefore results are subject to possible nonresponse bias.

A list of the participants is provided below with a description of limitations of the populations assessed. Unless noted, 90 percent or more of the age-eligible children in a population are in school. For countries where more than 10 percent of the age-eligible children are out of school a notation of *in-school population* appears after the country's name. In Brazil, two separate samples were drawn, one each for the cities of São Paulo and Fortaleza. In Canada, nine out of the 10 provinces drew separate samples of 13-year-olds and four of these drew separate samples of English speaking and French-speaking schools, for a total of 14 separate samples. Four Canadian provinces — six separate samples — participated in the assessment of 9-year-olds.[2] These distinct Canadian samples coincide with the separate provincial education systems in Canada and reflect their concern for the two language groups they serve. The IAEP project was asked to provide separate results for the American state of Colorado, which opted to assess its 9- and 13-year-olds students in mathematics, science, and geography.

[2] Taken together, the Canadian samples represent 94 percent of the 13-year-olds and 74 percent of the 9-year-olds in Canada. An appropriately weightd subsample of responses was drawn from these samples for the calculation of the statistics for Canada.

Participants

Brazil	Cities of São Paulo and Fortaleza, restricted grades, in-school population
Canada	Four provinces at age 9 and nine out of 10 provinces at age 13
China	20 out of 29 provinces and independant cities, restricted grades, in-school population
England	All students, low participation at ages 9 and 13
France	All students
Hungary	All students
Ireland	All students
Israel	Hebrew-speaking schools
Jordan	All students
Korea	All students
Mozambique	Cities of Maputo and Beira, in-school population, low participation
Portugal	Restricted grades, in-school population at age 13
Scotland	All students, low participation at age 9
Slovenia	All students
Soviet Union	14 out of 15 republics, Russian-speaking schools
Switzerland	15 out of 26 cantons
Taiwan	All students
United States	All students

Each participating country was responsible for carrying out all aspects of the project, including sampling, survey administration, quality control, and data entry using standardized procedures that were developed for the project. Several training manuals were developed for the IAEP project. These comprehensive documents, discussed with participants during several international training sessions, explained in detail each step of the assessment process.

Typically, a representative sample of 3300 students from 110 different schools was selected from each population at each age level and half were assessed in mathematics and half in science. A total of about 175,000 9 and 13-year-olds (those born in calendar years 1981 and 1977, respectively) were tested in 13 different languages in march 1991.

Initial results of the second IAEP have been reported in *Learning Mathematics* and *Learning Science*.[3] Reports of the geography and performance assessments were issued in June and July 1992, respectively.[4] Mathematics and science results for Colorado were reported in May 1992 and geography results, in August 1992.[5] A technical report published in April 1992 describes the tools that were used to obtain these results, which was considered to be the first stage of data analysis.[6]

In a second stage of data analysis, scales for mathematics and science proficiency were obtained using what can be called a strong model-based psychometric strategy, item response theory. Item response theory utilizes a family of models that employ latent variables (i.e., variables that cannot be observed) that correspond to the dimensions of what is known as the "latent space". As mentioned in the introduction, a technical report published in November 1992 gives a detailed description of the tools used in this second stage.[7]

The present paper focus on a part of the second stage of data analysis: the work done regarding the creation of a unique scale for all the participating populations, i.e. creating reference populations, scaling methodology, linkage of 9 and 13-year-olds populations.

[3] Archie E. Lapointe, Nancy A. Mead, and Janice M. Askew. *Learning Mathematics*. Princeton, NJ; Educational Testing Service, 1992.
Archie E. Lapointe, Nancy A. Mead, and Janice M. Askew. *Learning Science*. Princeton, NJ; Educational Testing Service, 1992.

[4] Stephen Lazer. *Learning About the World*. Princeton, NJ: Educational Testing Service, 1992.
Brian McLean Semple. Performance Assessment: An International Experiment. Edinburgh, Scotland: Scottish Education Department, 1992.

[5] Ruth B. Ekstrom. *Colorado: Meeting the Challenge in Mathematics and Science*. Denver, CO: Colorado Department of Education, 1992.
Ruth B. Ekstrom. *Colorado: Meeting the Challenge in Geography*. Dever, CO: Colorado Department of Education, 1992.

[6] Adam Chu, et al. *IAEP Technical Report*. Princeton, NJ: Educational Testing Service, 1992.

[7] Jean-Guy Blais, et al. *IAEP Technical Report: vol.2*. Princeton, NJ: Educational Testing Service, 1992.

Reference populations

The presence of so many different and heterogeneous populations (over 20 for 9-year-olds and over 30 for 13-year-olds) makes the task of creating a common scale an interesting technical and theoretical problem. How can we create a scale with which we can reasonably compare different populations? This could be accomplished in different ways using different reference populations. In this study, no single population was chosen to serve as the reference population. Instead, all populations contributed to creating four combined reference populations, called "superpopulations", one for each age group by mathematics and science combination.

All superpopulations were initially formed by drawing random samples of 200 students from each of the participating populations.[8] At age 9, 2,800 examinees were retained for each of the mathematics and science reference populations. At age 13, 4,000 and 3,800 examinees were retained for mathematics and science populations, respectively. All the analyses were conducted with these four data sets, but a number of analyses were repeated using the full populations. This was the case for item parameter estimation and plausible values computation. Analyses of the superpopulations were conducted without weights (i.e., each sampled student had a weight of one). Analyses of the full populations were conducted using transformed weights that summed to 1650. This transformation was necessary because some of the procedures could be affected by the number of examinees.

The analyses were based on the items retained after the first stage of the data analysis (see the first technical report). Sixty-one items were retained for 9-year-old mathematics and 75 items, for 13-year-old mathematics. Fifty-eight items were retained for 9-year-old science and 64 items, for 13-year-old science. For some populations, there were additional items that had to be deleted due to local problems in the translation of printed material. The maximum number of items removed was three.

[8] Canadian provinces did not contribute directly to the reference populations. Instead 200 examinees were randomly sampled from a population that had previously been labeled "Canada" (see the first IAEP technical report).

Scaling methodology

The Scaling Model

The paragraphs that follow review the sca'ing model employed in the analysis of the IAEP data. The reader is referred to Mislevy (1991) and Mislevy, Beaton, Kaplan & Sheehan (1992) for an introduction to plausible values methods and a comparison with standard psychometric analyses, to Mislevy, Johnson & Muraki (1992) for additional information, and to Rubin (1987) for the theoretical underpinnings of the approach.

The item response theory (IRT) scaling model used with IAEP is the 3-parameter logistic (3PL) model (e.g., see Lord, 1980). This model is from a family of "latent trait" models which quantify examinees' tendencies to provide responses in a given direction (e.g., correct answers), as a function of parameters that are not directly observed.

The fundamental equation of the 3PL is the probability that a person whose proficiency is characterized by the <u>unobservable</u> variable θ will respond correctly to item j:

$$P(X_j = 1 \mid \theta, a_j, b_j, c_j) = c_j + (1 - c_j) \frac{e^{Da_j(\theta - b_j)}}{1 + e^{Da_j(\theta - b_j)}}$$

$$\equiv P_j(\theta)$$

where: x_j is the response to item j, 1 if correct and 0 if not ;

a_j, $a_j > 0$, is the slope parameter of item j, characterizing its sensitivity to proficiency ;

b_j is the threshold parameter j, characterizing its difficulty ;

c_j , where $0 \leq c_j < 1$, is the lower asymptote parameter of item j, reflecting the chances of a correct response from students of very low proficiency.

In IAEP analyses, c parameters were estimated for multiple-choice items, but were fixed at zero for constructed response items.

Under the usual IRT assumption of local independence, the probability of a vector $\mathbf{x} = (x_1, ..., x_n)$ of responses to n items is simply the product of terms based on the fundamental equation of the 3PL:

$$P(\mathbf{x} \mid \theta, \mathbf{a}, \mathbf{b}, \mathbf{c}) = \prod_{j=1}^{n} \left[P_j(\theta) \right]^{x_j} \left[1 - P_j(\theta) \right]^{1 - x_j}$$

After x has been observed, this equation can be considered a likelihood function, and provides a basis for inference about θ or about item parameters. In IAEP, estimates of item parameters were obtained via a marginal maximum likelihood estimation procedure (see Bock & Aitkin, 1981) as implemented in Mislevy and Bock's (1982) BILOG computer program.

Overview of Plausible Values Methodology

A detailed development of plausible values methodology is given in Mislevy (1991). Along with theoretical justifications Mislevy's paper presents some secondary analyses and numerical examples. Plausible values were developed as a means of obtaining consistent estimates of selected population features, and approximations of others that are no worse than those that would be obtained using standard IRT procedures. The following paragraphs give a brief overview of the approach.

Let $\mathbf{y}$ represent the responses of all sampled examinees to background and attitude questions. If IRT θ values were available for all sampled examinees, it would be possible to compute a statistic $t(\boldsymbol{\theta}, \mathbf{y})$ -- such as a subpopulation sample mean, a sample percentile point, or a sample variance -- to estimate a corresponding population quantity T. A function $U(\boldsymbol{\theta}, \mathbf{y})$ -- e.g., a jackknife estimate -- would be used to gauge sampling uncertainty. Because the 3PL is a latent variable model, however, θ values are not observed even for sampled students. If enough responses are solicited from each student to provide a fairly precise estimate $\hat{\theta}$ of their θ values, values of $t(\hat{\boldsymbol{\theta}}, \mathbf{y})$ and $U(\hat{\boldsymbol{\theta}}, \mathbf{y})$ are reported as approximations of corresponding $t(\boldsymbol{\theta}, \mathbf{y})$ and $U(\boldsymbol{\theta}, \mathbf{y})$ values.

Following Rubin (1987), we can think of $\boldsymbol{\theta}$ as "missing data" and approximate $t(\boldsymbol{\theta}, \mathbf{y})$ by its expectation given $(\mathbf{x}, \mathbf{y})$, the data that were observed, as follows:

$$t^{*}(\mathbf{x}, \mathbf{y}) = E[\, t(\boldsymbol{\theta}, \mathbf{y}) \mid \mathbf{x},\ \mathbf{y}\,]$$
$$= \int t(\boldsymbol{\theta}, \mathbf{y})\, p(\boldsymbol{\theta} \mid \mathbf{x}, \mathbf{y})\, d\boldsymbol{\theta}$$

It is possible to approximate t^{*} by using random draws from the conditional distribution $p(\theta \mid x_i, y_i)$ of each sampled student i. These values are referred to as imputations in the sampling literature, and as "plausible values" in IAEP. The value of θ for any respondent that would enter in the computation of t is thus replaced by a randomly selected value from the conditional distribution for θ given his or her responses to cognitive items (x_i) and background items (y_i). Rubin (1988) proposes this process be carried out several times -- multiple imputations -- so that the uncertainty associated with imputation can be quantified. The average of the results of K estimates of t, each computed from a different set of plausible values, is a Monte Carlo approximation of the above integral; the variance among them, denoted by B, reflects uncertainty due to not observing θ, and must be added to the estimated expectation of $U(\boldsymbol{\theta}, \mathbf{y})$, which reflects uncertainty due to testing only a sample of students from the population.

Plausible values are not test scores for individuals in the usual sense. They are offered only as intermediary computations for calculating integrals of the form presented above in order to estimate population characteristics, even though they are biased estimates of the proficiencies of the individuals with whom they are associated.

Computing Plausible Values in IRT-based Scales

Plausible values for each respondent i are drawn from the conditional distribution $p(\theta \mid x_i, y_i)$. This section describes how, in IRT-based scales, these conditional distributions are characterized and how the draws are taken.

Using conditional independence we have:

$$p\,(\theta \mid x_i, y_i) \propto P(x_i \mid \theta)\, p(\theta \mid y_i)\ ,$$

where $P(x_i \mid \theta)$ is the likelihood function for θ induced by observing x_i (treating item parameter estimates as known true values) and $p(\theta \mid y_i)$ is the distribution of θ given the observed value y_i of background responses.

In the analysis of IAEP data, a normal (Gaussian) form was assumed for $p(\theta \mid y_i)$, with a common dispersion and with a mean given by a main-effects model for selected elements of the complete vector of background variables. The background variables included will be referred to as the conditioning variables[9], and will be denoted y^c. The following model was fit for each subject group for each age (i.e., mathematics and science for 9- and 13-year-olds):

$$\theta = \Gamma y^c + \varepsilon \quad ,$$

where ε is normally distributed with mean 0 and dispersion Σ; and Γ and Σ are the parameters to be estimated. Since the subject areas in IAEP were considered to have just one scale, Γ is a vector and Σ is a scalar. If we had decided to use subscales, then Γ would have been a matrix and Σ a covariance matrix. Like item parameter estimates, these estimates of conditional distributions were treated as known true values in subsequent steps of the analysis. Maximum likelihood estimates of Γ and Σ were obtained with Sheehan's (1985) M-Group computer program, using a variant of the EM solution described in Mislevy (1985).

The conditional distribution, $p(\theta \mid y_i)$, has been assumed normal, with mean $\mu_i^c = \Gamma y_i^c$ and variance Σ; if the likelihood, $P(x_i \mid \theta)$, is approximated by another normal distribution, with mean μ_i^L and variance Σ_i^L, then the posterior $p(\theta \mid x_i, y_i)$ is also normal with variance:

$$\Sigma_i^p = (\Sigma^{-1} + (\Sigma_i^L)^{-1})^{-1}$$

and mean: $$\theta_i = (\mu_i^c \, \Sigma^{-1} + \theta_i^L \, \Sigma_i^L) \, (\Sigma_i^p)^{-1} \quad .$$

In the IAEP analysis, a normal approximation for $P(x_j \mid \theta)$ was accomplished for a given scale by the steps described below. These computations were carried out in the scale determined by parameters estimates from different runs of BILOG (Mislevy & Bock, 1982).

1- Lay out a grid of Q equally spaced points from -5 to +5, a range that should cover the region of the scale for each population involved. The number of Q values varies from 20 to 40, depending on the scale being used; smaller number of values should suffice for scales with few items given to each respondent, while larger number of values are required for scales with many items (such as in IAEP).

[9] The conditioning variables used in IAEP analyses are presented in the second technical report, appendix C. The way they were included in the analysis is described in a further section.

2- At each point X_q, compute the likelihood $L(x_i \mid \theta = X_q)$.

3- To improve the normal approximation in those cases in which the likelihoods are not roughly symmetric in the range of interest -- as when all of an examinee's answers are correct -- multiply the values from step 2 by the mild smoothing function

$$S(X_q) = \frac{\exp(X_q + 5)}{[1 + \exp(X_q + 5)]\,[1 + \exp(X_q - 5)]}$$

This is equivalent to augmenting each examinee's response vector with responses to two fictitious items, one extraordinarily easy item that everyone gets right and one extraordinarily difficult item that everyone gets wrong. This expedient improves the normal approximation for examinees with flat or degenerate likelihoods in the range where their conditional distributions lie, but has negligible effects for examinees with even modestly well-determined symmetric likelihoods.

4- Compute the mean and standard deviation of θ using the weights $S(X_q)$ from step 3

At this stage, then, the likelihood created by a respondent's answers to the items in a given scale is approximated by a normal distribution. This normalized-likelihood normal posterior approximation is then employed in both the estimation of Γ and Σ and in the generation of plausible values. From the final estimates of Γ and Σ , an examinee's posterior distribution is obtained from the normal approximation using the four-step procedure outlined above and a plausible value is drawn at random from this univariate normal distribution.

Even though we do not observe the θ value of examinee i, we do observe variables that are related to it: x_i, the examinee's answers to the cognitive items, and y_i, the respondents answers to demographic and background variables. Suppose we wish to draw inferences about a number $T(\mathbf{\Theta}, \mathbf{Y})$ that could be calculated explicitly if the θ and y values of each member of the population were known. Suppose further that if θ values were observable, we would be able to estimate T from a sample of N pairs of θ and y values by the statistic $t(\boldsymbol{\theta}, \mathbf{y})$ [where $t(\boldsymbol{\theta}, \mathbf{y}) = (\theta_1, y_1, \ldots, \theta_N, y_N)$], and that we could estimate the variance in t around T due to sampling respondents by the function $U(\boldsymbol{\theta}, \mathbf{y})$. Given that observations consist of (x_i, y_i) rather than (θ_i, y_i), we can approximate t by its expected value conditional on $(\mathbf{x}, \mathbf{y})$, or (as previously seen):

$$
\begin{aligned}
t^*(\mathbf{x}, \mathbf{y}) &= E[\, t(\boldsymbol{\theta}, \mathbf{y}) \mid \mathbf{x}, \mathbf{y}\,] \\
&= \int t(\boldsymbol{\theta}, \mathbf{y})\, p(\boldsymbol{\theta} \mid \mathbf{x}, \mathbf{y})\, d\boldsymbol{\theta}
\end{aligned}
$$

It is possible to approximate t^* with random draws from the conditional distributions $p(\theta_i \mid x_i, y_i)$, which are obtained for all examinees by the method describe above. Let $\hat{\boldsymbol{\theta}}_m$ be the m^{th} such vector of the "plausible values," consisting of a value for the latent variable of each examinee. This vector is a plausible representation of what the true $\boldsymbol{\theta}$ vector might have been, had we been able to observe it. The following steps describe how an estimate of a scalar statistic $t(\boldsymbol{\theta}, \mathbf{y})$ and its sampling variance can be obtained from M (> 1) such sets of plausible values.[10]

1- Using each set of plausible values $\hat{\boldsymbol{\theta}}_m$ in turn, evaluate t as if the plausible values were true values of $\boldsymbol{\theta}$. Denote the results $\hat{t}_m$, for m = 1, ...,M.

2- Using a variance estimation procedure, compute the estimated sampling variance of $\hat{t}_m$ denoting the result U_m.

3- The final estimate of t is:

$$t^* = \sum_{m=1}^{M} \frac{\hat{t}_m}{M}$$

4- Compute the average sampling variance over the M sets of plausible values, to approximate uncertainty due to sampling respondents:

$$U^* = \sum_{m=1}^{M} \frac{U_m}{M}$$

5- Compute the variance among the M estimates $\hat{t}_m$, to approximate uncertainty due to not observing θ values from respondents:

$$B_M = \sum_{m=1}^{M} \frac{(\hat{t}_m - t^*)^2}{M-1} \quad .$$

[10] Five sets of plausible values were used in each IAEP analysis and are provided on the IAEP public-use data tapes for secondary analysis.

6- The final estimate of the variance of t^* is the sum of two components:

$$V = U^* + (1 + M^{-1})\, B_M \quad .$$

The first term, U^* is related to the sampling variancė and, in IAEP, it is estimated through a resampling procedure call the jackknife. Briefly, we can say that this procedure computes a statistic with the full data set and computes the same statistic a certain number of times, taking into account the sampling plan, deleting each time some of the data and replacing them with contiguous data (thus creating so-called pseudo data sets). The standard error is then estimated using the sum of the squared differences between estimates with the full data set and estimates with the pseudo data sets. The second term, $(1 + M^{-1})\, B_M$, is related to measurement error. It is the estimate of the uncertainty due to not observing θ. It is computed as the sum of the squared differences between the mean using each plausible value and the mean of the plausible values means. These two components are combined as shown above to form a more realistic standard error estimate.

Estimating variability requires computing a statistic 165 times, including 33 computer runs to obtain an estimate and a variance estimate from each of the five sets of plausible values used in IAEP analyses. Because the cost of the full procedure was prohibitive, an approximate procedure was used to produce reasonable estimates at lower costs. We estimated t on each pseudo-data sets (in order to estimate variability due to tl latency of proficiency) but computed its jackknife variance on only one pseudo data set to estimate sampling variability.

The invariance principle of IRT

Item response theory comes from the work of Ferguson (1942), Lawley (1943), Lord (1952) and Rasch (1960). This modelization proposal rests on the hypothesis that there exists a relation between the probability of obtaining the observed result for a given item and the non-observable ability (one or many) aimed at in the measurement process devised. It is inspired by statistical regression and by factor analysis. It supposes the existence of non-observable elements, traits, abilities, factors, influencing performance and observed results.

Modelizing in the statistical regression framework, as we find it in I.R.T., gives a theoretical property to the mathematical model's parameter: the invariance property. Under some conditions, estimates of items' parameters are independent of the group of examinees with which the measurement is done, and examinees' ability is independent of the items included in the measurement process (see Blais & Ajar, 1992; Blais & Ajar, 1993).

The invariance property is almost mythical because its empirical demonstration is remarkebly lacking in published studies, rendering it suspect on that account (Wood, 1976). There's been a lot of confusion about it in I.R.T.'s applications and Wood (1976) said that it was one of the grayest concept in general test theory. In some presentation of the invariance property, it is suggested that the model guarantees invariant estimation, that invariant estimations can be obtained with any group of examinees or items without empirical investigations. As if I.R.T. model spared us from worrying about anything any more. Fortunately of course, reality is somewhat more complex. Most of the time there's far to go before counting the chickens.

Before getting into conditions for the invariance property to hold and into discussion of elements that could tamper it, we must place things in a general context. To do so, we will expose the idea of statistical regression as we find it in some statistics textbooks (for example, Cramer 1946, p. 270-272), the special case of linear regression and how it can be formulated in the framework of I.R.T.

Let X and Y be two continuous random variables and $f(x,y)$ their joint probability function. If we think of Y as a dependent variable and of X as an independent variable, then we can write $f(y|x)$, the probability function of y given x.

For a given value of X, say x, the Y variable can take many values y. A possible representative of these values could be the expected mean of Y given the value x taken by X: $E(Y|X=x) = \mu_{y|x}$.

When x varies, the point $[x, E(Y|X=x)]$ describes a curve in a two-dimensional space. A curve like this is called a regression curve, it is said to represent the regression of Y on X.

The regression of Y on X is independent of the x's distribution, it is invariant from one group of x values to any other one.

If we suppose a linear relationship between Y and X, we can represent it by the equation $Y = aX + b + e$, where e is an *error* variable. If Y and e are random variables and $E(e)=0$, then $E(Y|X) = aX + b$, i.e. the regression of Y on X is given by $aX + b$.

When for estimated values of a and b, the *hypothesis of a linear relationship can be confirmed*, the established relation will be the same whatever values of X are considered. In other words the invariance property will hold for the a and b parameters.

The parallel with I.R.T.'s modelization can be illustrated with the case where the notation is dichotomous.

Let the result (0 or 1) to a given item j be the variable U_j. $P_j(q)$ can be the probability of a correct result, noted 1, given an ability q: $P_j(q) = P(U_j=1|q)$. And $Q_j(q)$ the probability of an incorrect result, noted 0, given an ability q: $Q_j(q) = P(U_j=0|q) = 1 - P(U_j=1|q)$.

Let's suppose that the probability function of U_j is of the Bernouilli type, then:

$$f_j(U_j) = \begin{cases} P_j(\theta), u_j = 1 \\ Q_j(\theta), u_j = 0 \end{cases}$$

The regression of the observed result, for item j, on the ability q is given by:

$$E(U_j|q) = [\ P_j(q) \times 1\ ] + [\ Q_j(q) \times 0\ ]$$

$$= P_j(q).$$

We called the regression of the observed result U_j, for item j, on the ability q, the characteristic curve of item j: **I.C.C.** The characteristic curve is invariant for any distribution of the ability variable q. If $P_j(q)$ is in the form of a two-parameter logistic model, i.e. one parameter q for the examinees and two parameters (a, b) for the items, since the I.C.C. is invariant the parameters are, theoretically, also invariant.

For the property to hold empirically, the modelization must meet some conditions. Certainly, an important one is that there must be some appropriateness between the data and the model. Like for any regression model, the invariance property holds only if we can demonstrate that the model fits the data reasonably well. The main difficulty resides in defining what is meant by *reasonably well.*

From the modelization point of view, goodness of fit is not the only concern that must get the practitioner's attention, even if it is an important point. At the initial calibration of items, the road that leads to parameters' estimation for our model, we must have some thought about the desired level of generalization we are aiming at for our results. In direct link with our measurement objectives we must have considerations for the way we sample the examinees. Not that we must use a specific probabilistic sampling plan -we are in the regression framework- but the sample must be conveniently heterogeneous (Hambleton and Swaminathan, 1985, p. 13). The task we have to tangle with is to give a practical meaning to the word *conveniently.*

In a new measurement situation, with items previously calibrated, invariance will hold only if the new examinees being tested have similar characteristics as those in the group that was useu for the initial calibration of items (Lord & Novick, 1968 p. 360). We must then take into consideration the heterogeneity wanted in the initial calibration if we want to collect useful information that will help us explore the promise of invariance with future examinees' samples.

When there is an interaction between a group of examinees and some given items, so that items have a different meaning for different groups (a sort of bias), the invariance property will not hold any more.

Scaling

First, let us review the entities we were dealing with in this scaling analysis. As described in a previous section, three elements contribute to a population's mean proficiency estimate: supposedly known item parameters, known answers to background items, and known answers to cognitive items. Thus, for each analysis, there were three data sets to be considered for each subject area and age group.

The information that is known can be used directly. The content of the cognitive items is documented in the first IAEP results report published in February 1992. All the cognitive items included in the percents correct analyses in the first stage of data analysis were used for the IRT analyses.

The background variables used for computing plausible values were not used directly. Since we are dealing almost exclusively with qualitative variables, and since plausible values methodology can be seen as a kind of regression analysis, all background variables were transformed into "dummy" variables. These variables took the form of a series of orthogonal comparisons which described the various categories of each background variable. These variables

Figures 7 and 8 show fitted regression lines of the two proficiency scores. (Individual population proficiency scores were regressed against the reference population proficiency scores.) The graphics presented at the top and bottom of Figure 7 are for mathematics 9- and 13-year-olds, respectively. The graphics presented at the top and bottom of Figure 8 are for science 9- and 13-year-olds, respectively.

Insert figures 7 and 8 about here

These last figures indicate that the proficiency distributions estimated from the reference populations cannot be considered as having the same mean and standard deviation as the distributions estimated from the individual populations. The proficiency estimates are highly related, with correlations of over 0.98, but the midpoint and scale of the proficiency distributions are different. The midpoint and scale of a proficiency distribution is arbitrary. To compare the proficiency distribution directly from population to population, they must be put onto the same metric. This is accomplished by equating.

The scales that result from separate IRT scalings are not typically comparable, even if the same set of items are used in each of the scalings. The origin and scale units of the provisional scale for each of the individual population's scaling and for the reference population's scaling were established by setting the ability distribution of each of the respective calibration samples to a mean of zero and a standard deviation of one. Because all participating populations were used in the reference population's item calibrations, the origin and scale unit for the reference populations were based on the sum of the individual populations' ability distributions. In contrast, the origin and scale unit of the individual populations' scales were based on an ability distribution of each single population. Clearly, without a transformation, the metrics for the individual populations are not comparable to one other or to the metrics of the reference populations. Consequently, additional procedures were employed to ensure all scores were reported on the same metric.

The next step then was to put the item parameters estimates coming from the individual populations and those coming from the reference population for a given subject area and age group on the same metric. This was done by equating the item parameters estimated from the individual populations to those estimated from the reference populations (see Section for documentation of the equating procedures). After the equating was completed, proficiency scores were recalculated using equated estimated item parameters and compared to proficiency scores computed from item parameters estimated from the reference populations.

Figures 9 and 10 illustrate what happened to previous regression lines when equated item parameters estimates were used instead of those that were not equated. The lines are much more homogeneous and, except for one population, 13-year-old mathematics, we could say with confidence that equated item parameters estimates and those coming from the reference populations were on the same metric. Moreover, the rank orders of the mean proficiency scores for populations based on the equated ICCs and those based on the reference population ICCs are the same (if standard errors are taken into account). These results indicate that no information has been lost as a result of using one set of item parameter estimates over another. We, therefore, decided to use the overall estimates based on the reference populations and five plausible values were computed for each examinee from each population using Sheehan's M-Group program (the mainframe version).

Insert figures 9 and 10 about here

To assess the relationship between proficiency scores and previously presented percent correct scores, correlations were computed between each plausible value and mean percent correct score for each population. The results are presented in Tables 3 to 6. As we can see from these tables, the correlation between the mean percent correct and the mean of the plausible values (column labeled CORR) and each individual plausible value (columns labeled P1 to P5) were quite high. We can also observe that the rank order based on mean percent correct scores (column labeled %) corresponds to rank order based on mean proficiency scores (column labeled PROF). The mean proficiency scores are presented on a scale with a mean of 500 and a standard deviation of 100. The next paragraphs will describe the transformation applied to the original proficiency scale (≈ [-3.00, 3.00]).

Insert tables 3 to 6 about here

To be able to perform a linear transformation of an existing scale, one has to know what are its initial mean and standard deviation and the targeted mean and standard deviation. For IAEP, the target mean and standard deviation were fixed at 500 and 100 respectively. These values were chosen mainly for reporting convenience.

The initial mean and standard deviation had to be calculated from the existing data. In creating a common scale using all the participants we decided that the initial mean and standard deviation would be calculated using the weights of each examinee in each populations for a subject area. After equating the results of the 9- and 13-year-old populations for a given subject area (mathematics or science)[11], all the examinees in these two age-groups were put together and initial values were calculated. The following summarizes how this was done.

Let $\overline{X}_{\theta_1}$ and $S^2_{\theta_1}$ stand for the mean and standard deviation of the initial scale and $\overline{X}_{\theta_2}$ and $S^2_{\theta_2}$ stand for the mean and standard deviation of the targeted scale, respectively. Then θ_1, a plausible value on the initial scale, is transformed to θ_2 , a plausible value on the transformed scale, by calculating:

$$\theta_2 = \left[\frac{S_{\theta_2}}{S_{\theta_1}} (\theta_1 - \overline{X}_{\theta_1})\right] + 500 \quad .$$

Finally, the mean proficiency score was calculated for each population. Tables 8 and 9 present some of the results.

Column one is a population ID. (The asterisk beside the standard error indicates a 9-year-old population.) Column two gives the mean proficiency score for each population. Column three gives the standard error of the mean proficiency score. Remember that this standard error takes into account sampling and imprecision in measurement. The next eleven columns gives information on the distribution of scores. They provide the first, tenth, twentieth, thirtieth, fortieth, fiftieth, sixtieth, seventieth, eightieth, ninetieth, hundredth percentiles of the distribution, respectively for each population.

Insert tables 8 and 9 about here

Linking 9- and 13-year-olds populations

A small set of common items were included in the 9- and 13-year-olds assessments in each subject area. (Fourteen of these common items in each subject were retained after the first stage of data analysis.) This design element provided the possibility of linking 9- and 13-year-olds results into a single scale within each subject area (i.e., mathematics and science).

[11] This part of the analysis is documented in the following section.

The item parameters of the common items were estimated independently using the 9-year-old reference population and 13-year-old reference population. These item parameter estimates were different because the scales defined by each independent calibration of the items were different. In order to merge the two age group proficiency scores, we had to make sure they were expressed on the same scale. There are a number of methods for transforming item parameter estimates from one scale to another scale (see Stocking & Lord, 1983).

Two procedures for transforming IRT results to a common scale are common items equating and equivalent populations equating. The common items equating procedure is used to equate two scales that contain a set of common items that were administered to independent samples drawn from different populations. This was the case for the 9- and 13-year-old IAEP populations.

The procedure that was used to estimate the transfo.mation for linking the two age groups was the Stocking-Lord procedure (Stocking & Lord, 1983) as implemented in the TBLT computer program (Stocking, 1986).

The input data for the Stocking-Lord procedure consists of two sets of estimated item parameters, one set expressed on a target scale and one set expressed on a provisional scale. In the IAEP study the 13-year-old scale was chosen as the target scale and the 9-year-old scale as the provisional scale. The output of the Stocking-Lord procedure are the parameter estimates, denoted here by A and B, based on a linear transformation that describes the relationship between the IRT item parameter estimates expressed on the provisional scale and those expressed on the target scale. That is,

$$a_j = A^{-1} a_j^p$$
$$b_j = A\, b_j^p + B$$
$$c_j = c_j^p$$

where (a_j^p, b_j^p, c_j^p) and (a_j, b_j, c_j) for $j = 1,\dots,n$ are IRT parameter estimates obtained for the common items expressed on the provisional and target scales, respectively. Note that the lower asymptote parameters c_j^p are unaffected by the transformation.

The parameters of the linear transformation, A and B, are found by minimizing the squared difference between estimated true scores (expected numbers correct on the n common items) at N preselected proficiency values, $\theta = [\theta_1,...,\theta_N]$. The function that is minimized is :

$$f(A, L) = 1 / N \sum_{i=1}^{N} \left\{\zeta^T(1, 0, \theta_i) - \zeta^P(A, B, \theta_i) \right\}$$

where $\zeta^T(1, 0, \theta_i)$ is the estimated true score associated with the proficiency level θ_i, calculated from the item parameters expressed on the target scale, and $\zeta^P(A, B, \theta_i)$ is the estimated true score associated with the proficiency level θ_i calculated from the item parameters that were originally estimated on the provisional scale and then re-expressed on the target scale. That is,

$$\zeta^P(A, B, \theta_i) = \sum_{j=1}^{n} c_j + \frac{(1 - c_j)}{\left\{1 + \exp\left[-1.7(A^{-1}a_j^P)(\theta_i - (Ab_j^P + B))\right]\right\}}$$

where a_j^P and b_j^P are the estimated discrimination and difficulty parameters for item j, expressed on the provisional scale. The values $\theta = [\theta_1,...,\theta_N]$ are typical selected to span that region of the target scale which is expected to be the most dense.

The transformations were obtained using Stocking (1986) TBLT program. The equated item parameters estimates are presented in Tables 9 to 14.

Insert tables 9 to 14 about here

The transformation used for putting 9-year-olds (b_j^P) item parameter estimates on to the 13-year-olds scale, for mathematics was

TARGET = (1.076767 x PROVISIONAL) + (-1.323902),

and for science,

TARGET = (1.116443 x PROVISIONAL) + (-1.350289).

Of course, obtaining such equations does not mean that estimates are identical. The quality of the linking procedure must still be checked.

For mathematics, we can see by looking at Tables 9 to 11 and Figures 11 to 13 (where 9-year-old estimates are on the horizontal axis and 13-year-olds are on the vertical axis) that the main problem lies with the difficulty parameter of one item, item 30. In Figure 12, this item is the point at the extreme right (0.43, -0.62). Otherwise, even if values do not fall exactly on a straight line, we can consider that parameters at each age are similar. (Remember that the c values were not equated.) Correlation between estimates of the a and b parameters were 0.67 and 0.80, respectively.

Insert figures 11 to 13 about here

For science, we can see by looking at Tables 12 to 14 and Figures 14 to 16 there are two items which present a problem, item 26 and item 27. These are the same items that are singled in the next chapter on the item anchoring procedure. Correlation between estimates of a and b parameters were 0.67 and 0.73, respectively.

Insert figures 14 to 16 about here

What should be done with these items? Ideally they should be removed and the equating redone. However, the number of common items in the IAEP assessment were limited, and the performance of the linking procedure is at least partly affected by the number of items included in the linking. This is because the procedure is sensitive to uncertainty due to model misfit, which becomes more severe as the number of linking items decreases (Sheehan, 1988). In a sense, as the number of items increases, the effect of a few peculiar items should decrease. Because we felt that mean proficiency score estimates would be robust to the presence a small number of non "ideal" items in the equating procedure and we wanted to keep as many items as possible (content coverage being important), the decision made was to keep all the items and go ahead with the linking of the 9- and 13-year-olds scales, keeping in mind that there is always some uncertainty due to any linking procedures (as well illustrated by Sheehan, 1988).

Conclusion

Results show that even if on an item by item basis the equating gives results that are not "ideal", when all the items are taken into account (i.e. the test format) the students proficiency scores estimated with both sets of item parameters estimate (from the reference populations and from the "equated" individual populations) can be considered to be on the same scale. Moreover, the rank orders of the mean proficiency scores for populations based on the equated ICCs and those based on the reference population ICCs are the same (if standard errors are taken into account). These results indicate that no information has been lost as result of using one set of item parameter estimates over another.

Implications for applied research and educational data analysis could be important. When working with multiple heterogeneous populations, one can choose to work with a reference population that is a mixture of individual populations. In doing so, there should be no loss of information if one is working on an aggregated measures basis (test scores or population' mean test scores). However, one has to be careful when working with individual items (like in computerized adaptive testing), since there could be some important discrepancies in between individual population equated item parameters estimate.

References

Blais, J.-G. & Ajar, D. (1992). Théorie des réponses aux items et modélisation. *Mesure et évaluation en éducation*, **14**, 5-18.

Blais, J.-G. & Ajar D. (1993). Modelizing test scores with item response theory: It is not one size fits all. Paper submitted for publication to the Journal of Educational Measurement.

Cramer, H. (1942). *Mathematical Models of Statistics*. Princeton: Princeton University Press.

Ferguson, G.A. (1942). Item selection by the constant process. *Psychometrika*, **7**, 19-29.

Hambleton, R. K. & Swaminathan, H. (1985). *Item Response Theory: Principles and Applications*. Boston: Kluwer-Nijhoff.

Lawley, D. N. (1943). On problems connected with item selection and test construction. *Proceedings of the Royal Society of Edinburgh*, **61**, 273-287.

Lord, F. M. (1952). A theory of test scores. *Psychometrika Monograph*, No. 7.

Lord, F. M. (1980). *Applications of Item Response Theory to Practical Testing Problems*. New Jersey: Erlbaum Associates.

Lord, F. M. & Novick M. R. (1968). Statistical Theories of Menta. Test Scores. Reading, Mass.: Addison-Wesley.

Mislevy, R.J. (1985). Estimation of latent group effects. *Journal of the American Statistical Association*, **80**, 993-997.

Mislevy, R.J. (1991). Randomisation-based inference about latent variables from complex samples. *Psychometrika*, **56**, 177-196.

Mislevy, R.J., Beaton, A.E., Kaplan, B. & Sheehan, K.M. (1992). Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement*, **29**, 133-161.

Mislevy, R.J. & Bock, R.D. (1990). *BILOG 3: Item analysis and test scoring with binary logistic models* (Computer program). Chicago: Scientific Software.

Mislevy, R.J., Johnson, E.J. & Muraki, E. (1992). Scaling procedure in the National Assessment for Educational Progress. *Journal of Educational Statistics*, **17**, 133-161.

Nelson, J. (1987). *SWEEP* (Computer program). Princeton, NJ: Educational Testing Service.

Rasch, G. (1960). *Probabilistic Models for some Intelligence and Attainment Tests.* Copenhagen: Danish Institute for Educational Research.

Rubin, D.B. (1987). *Multiple imputation for nonresponse in surveys.* New York: John Wiley &Sons.

Sheehan, K.M. (1985). *M-GROUP: Estimation of group effects in multivariate models* (Computer program). Princeton, NJ: Educational Testing Service.

Sheehan, K.M. (1988). The IRT linking procedures used to place the 1986 intermediary scaling results onto the 1984 reading calibrating scale. In A.E. Beaton, *Expanding the new design: The 1985-1986 technical report* . (No. 17-TR-20) Princeton, NJ: National Assessment of Educational Progress.

Stocking, M.L. (1986). *TBLT: Transforming a's and b's by using a least squares technique* (Computer program). Princeton, NJ: Educational Testing Service.

Wood, R. (1976). Trait measurement and item banks. In D.N.M. de Gruitjer & L.J.Th. van der Kamp (Eds), *Advances in Psychological and Educational Measurement*. New York: Wiley.

Tables and figures

Table 1 **Mathematics, age 9, items' parameters estimates and standard errors**

ITEM ID	ITEM LABEL	SLOPE	SE	TRESHOLD	SE	GUESSING	SE
1-1M1	NUM-PK	1.15358	0.06	-1.17706	0.0702	0.24847	0.03633
2-1M1	MEA-PS	0.68846	0.03551	-1.61817	0.12152	0.21652	0.0467
3-1M1	NUM-CU	0.83867	0.04437	-0.85243	0.08628	0.22833	0.03664
4-1M1	NUM-CU	0.66089	0.063	0.08477	0.13	0.43187	0.03356
5-1M1	ALG-PS	0.5118	0.03437	-2.01419	0.20448	0.27147	0.05706
6-1M1	MEA-CU	0.46526	0.02995	-1.67893	0.19603	0.23938	0.05257
7-1M1	DAT-PK	0.91269	0.04814	-0.135	0.05778	0.20379	0.02479
8-1M1	NUM-PK	0.66594	0.02626	-1.64992	0.05606	0	0
9-1M1	MEA-CU	0.70783	0.02592	-1.6543	0.05465	0	0
10-1M1	GEO-CU	0.71741	0.02396	-0.74371	0.0298	0	0
11-1M1	ALG-CU	0.79017	0.02491	-0.40392	0.02356	0	0
12-1M1	NUM-PK	1.1168	0.03319	-0.37642	0.0185	0	0
13-1M1	NUM-PK	0.70757	0.02339	-0.17404	0.02433	0	0
14-1M1	MEA-PK	1.28359	0.069	0.31246	0.03136	0.19139	0.01543
15-1M1	NUM-PS	1.0947	0.06058	-0.16201	0.05055	0.22247	0.02377
16-1M1	NUM-CU	1.62095	0.08478	0.27242	0.02615	0.20983	0.01395
1-12M2A	NUM-PS	0.828	0.04876	-0.03503	0.06767	0.23484	0.02624
2-12M2A	NUM-PS	0.75785	0.04488	-0.45259	0.09108	0.25686	0.03387
3-12M2A	NUM-PS	0.8293	0.06614	0.78449	0.05577	0.27117	0.01938
4-12M2A	ALG-PS	0.92789	0.07725	0.99986	0.04764	0.28042	0.01586
5-12M2A	NUM-PK	0.77349	0.05877	0.52315	0.06707	0.27469	0.02324
6-12M2A	DAT-PS	0.97326	0.04868	-0.00624	0.04653	0.16562	0.02114
7-12M2A	NUM-CU	0.94739	0.05467	0.3143	0.0455	0.19152	0.01931
8-12M2A	NUM-PS	1.25314	0.10434	1.07732	0.03749	0.30381	0.01204
9-12M2A	DAT-CU	0.85102	0.05872	0.80681	0.04489	0.17341	0.01728
10-12M2A	NUM-PK	1.03049	0.03007	0.44814	0.01837	0	0
11-12M2A	NUM-CU	0.94808	0.12887	1.63286	0.07364	0.32057	0.01369
12-12M2A	NUM-CU	1.23887	0.10654	1.17726	0.03823	0.27199	0.01191
13-12M2A	ALG-PS	1.03295	0.06407	0.78132	0.03628	0.1609	0.01424
14-12M2A	GEO-CU	1.03056	0.0816	1.01171	0.04263	0.23691	0.01471
1-1M3	GEO-CU	0.58991	0.03844	-2.339	0.18846	0.26535	0.05722
2-1M3	DAT-CU	0.96807	0.0615	-1.03282	0.10084	0.38174	0.04066
3-1M3	DAT-CU	1.29945	0.07238	-0.61351	0.05688	0.32353	0.02812
4-1M3	NUM-PK	1.49841	0.06371	-0.73239	0.03849	0.16936	0.02369
5-1M3	MEA-PK	0.57301	0.039	-0.76622	0.14918	0.27294	0.04525
6-1M3	MEA-CU	0.73064	0.05218	0.01566	0.09242	0.31589	0.03014
7-1M3	NUM-PK	0.81144	0.04285	-0.34818	0.07109	0.19901	0.02894
8-1M3	NUM-CU	0.67596	0.02275	-0.46493	0.02765	0	0
9-1M3	NUM-PK	0.84861	0.02726	-0.56019	0.02591	0	0
11-1M3	GEO-PS	0.84981	0.02653	0.59891	0.0225	0	0
12-1M3	ALG-PS	1.00732	0.08466	0.32661	0.06365	0.41211	0.02206
13-1M3	NUM-CU	1.22822	0.07545	0.29677	0.03945	0.26364	0.01799
14-1M3	NUM-PK	0.892	0.04846	-0.40045	0.07199	0.22669	0.0309
15-1M3	NUM-CU	0.98494	0.05529	0.33756	0.04311	0.16772	0.01918
16-1M3	NUM-PK	1.5501	0.08653	0.40753	0.02678	0.2001	0.01371
1-1M4	GEO-CU	0.93718	0.05464	-1.50959	0.10792	0.32588	0.04636
2-1M4	NUM-PK	1.14108	0.05813	-0.7455	0.05883	0.21197	0.03083
3-1M4	NUM-PK	1.22622	0.05303	-0.62611	0.04716	0.19931	0.02537
4-1M4	DAR-PK	0.9241	0.05353	-1.07375	0.09911	0.35635	0.04004
5-1M4	DAT-PS	0.98672	0.05838	0.37687	0.04241	0.16773	0.0187
6-1M4	DAT-PK	1.16635	0.0583	-0.44396	0.05115	0.23954	0.02506
7-1M4	MEA-PK	0.62855	0.04601	-0.37126	0.12934	0.30776	0.03942
8-1M4	NUM-PK	1.11387	0.05929	-0.07947	0.04527	0.22027	0.02104
9-1M4	ALG-CU	0.34832	0.01831	-1.73093	0.09605	0	0
10-1M4	NUM-PK	0.60655	0.0216	-0.76071	0.03448	0	0
11-1M4	NUM-PS	0.90445	0.02645	-0.1567	0.0194	0	0
12-1M4	NUM-CU	1.04991	0.03073	0.50936	0.01824	0	0
13-1M4	MEA-PK	0.86948	0.02556	0.45126	0.02103	0	0
14-1M4	MEA-CU	0.91701	0.02727	0.24468	0.01881	0	0
15-1M4	NUM-PS	1.34951	0.07053	0.16408	0.03212	0.20245	0.01628
16-1M4	GEO-CU	0.95731	0.06168	0.52416	0.04504	0.20675	0.01847

Table 2 **Mathematics, age 13, items' parameters estimates and standard errors**

ITEM ID	ITEM LABEL	SLOPE	SE	TRESHOLD	SE	GUESSING	SE
1-2M1	GEO-CU	0.53152	0.02701	-1.03639	0.14854	0.21089	0.04779
2-2M1	ALG-PK	1.52416	0.05035	-0.31848	0.02463	0.16824	0.01364
3-2M1	NUM-CU	0.77062	0.03108	-0.78654	0.07639	0.17051	0.03288
4-2M1	MEA-CU	0.66237	0.03307	-0.16034	0.07958	0.19393	0.02778
5-2M1	NUM-CU	0.59752	0.03487	-0.36151	0.11255	0.22227	0.03656
6-2M1	NUM-PS	0.70937	0.03162	-0.96756	0.09884	0.18841	0.04108
7-2M1	NUM-PS	1.17333	0.05522	0.69702	0.02545	0.22404	0.0102
8-2M1	NUM-CU	1.39448	0.08895	1.17337	0.02702	0.31621	0.00799
9-2M1	MEA-PS	1.2339	0.04137	0.43553	0.01877	0.07072	0.00841
10-2M1	DAT-PK	0.54627	0.01585	-1.15737	0.03524	0	0
11-2M1	DAT-PK	0.96275	0.02022	-0.14431	0.01472	0	0
12-2M1	DAT-PK	0.8056	0.01767	-0.30767	0.01771	0	0
13-2M1	GEO-CU	1.06263	0.02194	-0.09168	0.01364	0	0
14-2M1	ALG-PS	0.91098	0.02172	0.84579	0.01836	0	0
15-2M1	ALG-PK	1.60427	0.05707	-0.33532	0.02445	0.15037	0.01459
16-2M1	DAT-PK	1.26213	0.04681	0.06054	0.02739	0.16437	0.01374
17-2M1	NUM-PK	1.18523	0.05278	0.09199	0.03303	0.19922	0.01605
18-2M1	MEA-PS	1.47208	0.06709	0.50954	0.02251	0.21778	0.01072
19-2M1	MES-PK	1.80225	0.09874	1.13716	0.01923	0.19567	0.00683
1-12M2B	NUM-PS	0.73187	0.02993	-1.44386	0.10678	0.20564	0.0476
2-12M2B	NUM-PS	0.83575	0.03072	-1.66592	0.08896	0.18289	0.04655
3-12M2B	NUM-PS	0.66709	0.0432	-0.14352	0.10024	0.3104	0.03102
4-12M2B	ALG-PS	0.59672	0.03163	-0.41038	0.10368	0.18696	0.03493
5-12M2B	NUM-PK	0.56451	0.03162	-0.56463	0.12584	0.20861	0.04049
6-12M2B	DAT-PS	0.67999	0.02609	-1.45374	0.10456	0.18137	0.04494
7-12M2B	NUM-CU	1.03708	0.05077	-0.58454	0.061	0.30164	0.02651
8-12M2B	NUM-PS	0.99587	0.03724	-0.38028	0.04429	0.17997	0.02025
9-12M2B	DAT-CU	0.82189	0.03403	-0.33349	0.05725	0.17542	0.02384
10-12M2B	NUM-PK	0.99254	0.02045	-0.74667	0.01791	0	0
11-12M2B	NUM-CU	1.11276	0.05668	-0.62607	0.06221	0.37623	0.02581
12-12M2B	NUM-CU	1.24063	0.04485	0.23263	0.02618	0.2069	0.01175
13-12M2B	ALG-PS	0.87039	0.04242	-0.60479	0.07411	0.24596	0.03094
14-12M2B	GEO-CU	1.07588	0.0475	-0.53731	0.053	0.24811	0.02488
15-12M2B	MEA-PS	1.2911	0.04336	0.35917	0.02007	0.09908	0.00954
16-12M2B	ALG-CU	1.30808	0.04722	0.16811	0.02501	0.1734	0.01218
17-12M2B	GEO-PS	1.04501	0.04233	0.06249	0.03601	0.19126	0.01607
18-12M2B	ALG-PK	1.28474	0.05651	0.5887	0.02388	0.21077	0.01048
19-12M2B	NUM-CU	1.23464	0.0533	0.90127	0.02077	0.12741	0.00798
1-2M3	NUM-CU	0.9243	0.0488	-0.74871	0.08205	0.32415	0.03292
2-2M3	NUM-PK	0.74802	0.03835	-1.00432	0.1131	0.28807	0.04331
3-2M3	DAT-CU	0.4541	0.02393	-2.40838	0.20256	0.23125	0.06048
4-2M3	MEA-CU	1.07484	0.03087	-0.00394	0.02274	0.05363	0.01045
5-2M3	GEO-CU	0.64575	0.04069	-0.74695	0.13943	0.31931	0.0439
6-2M3	ALG-PK	1.42528	0.0478	0.31203	0.01867	0.12614	0.00895
7-2M3	GEO-PS	1.03709	0.0367	-0.03512	0.031	0.12964	0.01418
8-2M3	NUM-CU	1.00965	0.05687	0.75797	0.03161	0.24858	0.012
9-2M3	MEA-PS	1.52737	0.07601	0.27018	0.02708	0.35802	0.01154
10-2M3	ALG-PK	1.29564	0.0258	-0.32567	0.0123	0	0
11-2M3	MEA-PK	0.78811	0.01754	-0.11313	0.01647	0	0
13-2M3	MEA-PK	1.11645	0.02295	0.19065	0.01276	0	0
14-2M3	GEO-PK	1.11033	0.05179	-0.03915	0.04108	0.30152	0.01698
15-2M3	NUM-CU	1.3071	0.05039	0.16072	0.02434	0.17517	0.01187
16-2M3	ALG-PS	1.71041	0.05515	0.43638	0.01421	0.09764	0.00668
17-2M3	ALG-PS	0.85475	0.03878	0.10066	0.04749	0.18707	0.01901
18-2M3	MEA-PS	1.55746	0.07126	0.86763	0.01874	0.16945	0.00758
19-2M3	MEA-PS	1.73092	0.10385	1.00672	0.02147	0.2686	0.00802
1-2M4	DAT-CU	0.5586	0.0264	-1.45686	0.15091	0.21829	0.05256
2-2M4	NUM-PK	1.36225	0.04727	-0.81677	0.03814	0.19853	0.02174
3-2M4	NUM-CU	1.0719	0.0561	0.38773	0.03704	0.31409	0.01396
4-2M4	NUM-PK	1.07789	0.04287	0.04027	0.03283	0.16783	0.0151
5-2M4	GEO-PK	0.85501	0.03016	-1.16239	0.06939	0.14685	0.03458
6-2M4	NUM-CU	1.3016	0.04317	0.59519	0.01652	0.06197	0.00676
7-2M4	ALG-CU	1.62637	0.0731	0.7426	0.01913	0.25316	0.0078
8-2M4	DAT-PK	1.26523	0.04181	-0.50717	0.03285	0.16329	0.01716
9-2M4	MEA-CU	1.25752	0.06774	0.55854	0.031	0.37872	0.01117

10-2M4	ALG-PK	0.84411	0.02072	-1.30959	0.02684	0	0
11-2M4	GEO-CU	1.246	0.02495	-0.20851	0.01223	0	0
12-2M4	NUM-PK	0.53821	0.01469	0.19705	0.02275	0	0
13-2M4	GEO-PS	1.12126	0.02335	-0.33497	0.01398	0	0
14-2M4	ALG-PK	1.18849	0.02528	0.09036	0.01241	0	0
15-2M4	NUM-PK	1.13158	0.02457	0.55491	0.01392	0	0
16-2M4	GEO-PK	1.07095	0.02358	0.30392	0.01369	0	0
17-2M4	NUM-PS	1.07142	0.05001	0.43645	0.03227	0.21714	0.01353
18-2M4	MEA-PS	1.81036	0.09149	0.84193	0.0194	0.262	0.00814
19-2M4	ALG-PS	1.24096	0.05742	0.63267	0.02462	0.1774	0.01057

Table 3 Mathematics, age 9, correlations between percent correct scores and plausible values

POP	%	PROF	CORR	P1	P2	P3	P4	P5
1	61.6	409.8	0.97	0.94	0.94	0.94	0.94	0.94
2	60.2	405.2	0.98	0.94	0.94	0.94	0.94	0.94
3	58.6	400.4	0.98	0.94	0.95	0.94	0.95	0.95
4	68.3	434.6	0.97	0.94	0.94	0.94	0.94	0.94
5	58.9	399	0.98	0.95	0.95	0.95	0.95	0.95
6	64.3	421.1	0.98	0.94	0.95	0.95	0.95	0.94
7	67.6	433.3	0.97	0.93	0.94	0.94	0.94	0.94
8	74.8	462.6	0.97	0.93	0.93	0.94	0.94	0.94
9	59.5	401	0.97	0.94	0.94	0.94	0.95	0.94
10	56.5	390.4	0.97	0.94	0.94	0.94	0.94	0.94
11	54.3	382.9	0.97	0.93	0.93	0.93	0.93	0.93
12	56.4	388.7	0.98	0.94	0.95	0.95	0.95	0.94
13	62.3	411.8	0.98	0.95	0.95	0.95	0.95	0.95
14	64.5	421.8	0.98	0.94	0.94	0.94	0.94	0.94
15	65.1	426.2	0.98	0.95	0.95	0.95	0.95	0.95
16	56.2	381.5	0.97	0.92	0.93	0.93	0.92	0.93
17	65.6	427.2	0.97	0.94	0.94	0.94	0.94	0.94
18	61.1	407	0.97	0.94	0.95	0.94	0.94	0.95
19	68.2	437.4	0.97	0.94	0.95	0.95	0.94	0.94
20	57.3	391	0.98	0.95	0.95	0.95	0.95	0.95

Table 4 Mathematics, age 13, correlations between percent correct scores and plausible values

POP	%	PROF	CORR	P1	P2	P3	P4	P5
1	63.8	522.1	0.98	0.95	0.95	0.95	0.95	0.95
2	65.9	531.1	0.97	0.95	0.95	0.95	0.95	0.95
3	62.1	517.7	0.97	0.95	0.95	0.95	0.95	0.95
4	80.1	582.3	0.95	0.91	0.91	0.9	0.91	0.92
5	60.6	515	0.97	0.95	0.95	0.96	0.95	0.95
6	33.9	407.4	0.9	0.84	0.85	0.85	0.85	0.85
7	64.4	525.7	0.98	0.95	0.95	0.95	0.96	0.96
8	68.3	539	0.97	0.95	0.95	0.95	0.95	0.95
9	60.5	512.4	0.97	0.95	0.95	0.95	0.95	0.94
10	63.5	523.1	0.97	0.95	0.95	0.95	0.95	0.95
11	64.7	523.4	0.97	0.95	0.95	0.95	0.95	0.95
12	40.7	442.3	0.94	0.9	0.9	0.91	0.9	0.91
13	73.6	557.3	0.96	0.94	0.94	0.94	0.94	0.94
14	57.7	503.1	0.97	0.94	0.94	0.94	0.94	0.94
15	63.3	521.3	0.97	0.94	0.94	0.95	0.95	0.94
16	30.8	401.3	0.84	0.74	0.74	0.73	0.73	0.74
17	57.5	501.6	0.97	0.94	0.95	0.94	0.94	0.95

18	60.9	512.6	0.97	0.94	0.95	0.95	0.95	0.95
19	59	508.7	0.97	0.94	0.94	0.94	0.94	0.94
20	59.9	511.9	0.97	0.95	0.95	0.95	0.95	0.95
21	57.8	505.1	0.97	0.95	0.94	0.94	0.94	0.95
23	53.9	491.7	0.97	0.93	0.94	0.93	0.93	0.94
24	50.2	478.1	0.96	0.93	0.93	0.93	0.93	0.93
25	66.1	531.4	0.97	0.95	0.95	0.95	0.95	0.95
26	68.9	537.8	0.98	0.94	0.95	0.95	0.95	0.95
27	36.7	423.2	0.93	0.88	0.88	0.88	0.88	0.88
28	62	517.5	0.98	0.95	0.95	0.94	0.95	0.95
29	67.7	534.4	0.97	0.94	0.94	0.93	0.95	0.95
30	60.7	515.8	0.98	0.95	0.96	0.95	0.96	0.95
31	57.6	504.9	0.98	0.95	0.95	0.95	0.95	0.95
32	70.3	544.7	0.97	0.95	0.95	0.95	0.95	0.95
33	55.7	492.6	0.97	0.93	0.93	0.93	0.93	0.93
34	74.2	555	0.97	0.94	0.94	0.94	0.93	0.94
35	72.7	561.8	0.95	0.93	0.93	0.93	0.94	0.93
36	54.6	491.4	0.97	0.95	0.95	0.94	0.95	0.94

Table 5 **Science, age 9, correlations between percent correct scores and plausible values**

POP	%	PROF	CORR	P1	P2	P3	P4	P5
1	65.4	436.4	0.96	0.91	0.91	0.91	0.9	0.91
2	62.2	417.3	0.96	0.9	0.9	0.91	0.9	0.91
3	62.5	418.8	0.97	0.91	0.92	0.92	0.91	0.92
4	62	418.8	0.95	0.89	0.9	0.89	0.89	0.89
5	54.8	379.5	0.96	0.91	0.91	0.91	0.91	0.91
6	60.9	411.5	0.96	0.91	0.91	0.91	0.91	0.92
7	66.4	440.9	0.96	0.91	0.91	0.91	0.91	0.91
8	68.4	441.9	0.96	0.9	0.9	0.9	0.91	0.91
9	60.8	409.3	0.96	0.9	0.91	0.91	0.91	0.91
10	61.6	414.2	0.96	0.91	0.91	0.91	0.91	0.91
11	55.6	381.4	0.95	0.88	0.88	0.88	0.88	0.88
12	54	373.5	0.95	0.89	0.88	0.88	0.89	0.89
13	62.3	419	0.96	0.91	0.92	0.91	0.91	0.91
14	62.4	417.3	0.96	0.89	0.89	0.9	0.89	0.9
15	61	413.2	0.96	0.9	0.91	0.91	0.91	0.9
16	57.9	383.1	0.94	0.86	0.86	0.87	0.86	0.86
17	61.2	414.3	0.96	0.9	0.9	0.9	0.91	0.9
18	61	410.4	0.96	0.91	0.91	0.91	0.91	0.91
19	66.2	437.4	0.97	0.92	0.92	0.92	0.92	0.92
20	63.6	426.4	0.97	0.93	0.93	0.93	0.93	0.93

Table 6 **Mathematics, age 9, correlations between percent correct scores and plausible values**

POP	NB	%	PROF	CORR	P1	P2	P3	P4	P5
1	1460	73.9	539	0.96	0.92	0.92	0.92	0.92	0.92
2	1617	72.5	533.2	0.96	0.91	0.92	0.91	0.92	0.91
3	4980	68.9	517.3	0.96	0.92	0.92	0.92	0.92	0.92
4	1775	66.7	510	0.97	0.92	0.93	0.92	0.93	0.92
5	929	68	516.1	0.97	0.92	0.93	0.94	0.94	0.93
6	1505	48.3	407	0.93	0.86	0.86	0.86	0.87	0.87
7	1787	68.4	516	0.97	0.94	0.93	0.93	0.94	0.93
8	1623	73.3	538	0.96	0.93	0.93	0.93	0.93	0.93
9	1657	63.1	492.6	0.97	0.93	0.93	0.93	0.93	0.93
10	1584	69.4	518.2	0.97	0.93	0.93	0.93	0.93	0.93

11	1485	70.6	521	0.97	0.92	0.93	0.93	0.93	0.93
12	1588	57.5	454.9	0.96	0.92	0.92	0.91	0.92	0.91
13	1635	77.5	556	0.97	0.92	0.92	0.92	0.93	0.93
14	1672	68.7	514.6	0.97	0.93	0.93	0.93	0.93	0.93
15	666	66.5	506.2	0.97	0.91	0.92	0.92	0.93	0.92
16	1604	66.3	504.6	0.97	0.92	0.93	0.92	0.92	0.93
17	1656	63.2	493	0.96	0.92	0.92	0.92	0.92	0.92
18	1566	66.7	504.7	0.97	0.93	0.93	0.93	0.93	0.93
19	1542	69.1	516.3	0.97	0.93	0.93	0.93	0.93	0.93
20	1609	66.9	509.2	0.96	0.92	0.92	0.92	0.92	0.92
21	1434	60.4	479.7	0.96	0.91	0.91	0.91	0.92	0.91
22	1520	63.7	490	0.97	0.93	0.93	0.92	0.92	0.92
23	1416	69.5	519.6	0.97	0.92	0.93	0.92	0.92	0.92
24	1579	71.3	528.5	0.96	0.92	0.92	0.92	0.92	0.92
25	1469	52.5	434.5	0.95	0.89	0.9	0.9	0.89	0.89
26	1694	70.6	521.7	0.97	0.92	0.93	0.92	0.93	0.93
27	223	65	500.7	0.96	0.92	0.9	0.9	0.91	0.9
28	1584	67.4	513.6	0.97	0.93	0.94	0.93	0.93	0.93
29	1598	70.4	521.4	0.97	0.93	0.92	0.92	0.93	0.93
30	1839	70.9	525.1	0.97	0.92	0.92	0.92	0.92	0.92
31	1609	68.2	508.1	0.96	0.92	0.92	0.91	0.91	0.92
32	3653	75.6	549.5	0.96	0.91	0.91	0.91	0.91	0.91
33	1786	75.8	548.6	0.97	0.94	0.93	0.93	0.94	0.94
34	1404	67	504.1	0.97	0.93	0.93	0.93	0.93	0.93

Table 7 Mathematics, age 9 and 13, mean proficiency scores and percentile by population

MEAN	SE	C 1	C 10	C 20	C 30	C 40	MED	C 60	C 70	C 80	C 90	C 100
381.51	1.93*	191.95	302.95	329.34	351.09	368.26	386.74	399.66	413.62	431.31	453.7	519.27
383.06	2.15*	175.97	298.51	329.01	353.24	372.77	388.7	402.94	417.25	435.31	457.95	522.98
388.03	3.92*	170.03	291.46	323.93	350.9	372.47	393.65	411.75	429.11	450.92	477.2	553.19
390.39	2.66*	164.18	296.37	332.78	358.03	378.49	397.15	412.85	429.43	448.72	474.92	547.83
391.26	4.38*	164.71	271.8	321.93	351.36	377.71	399.38	416.93	435.41	459.57	486.67	578.95
399.46	3.13*	171.11	289.46	337.99	366.86	387.93	408.49	423.98	442.74	463.34	489.78	585.16
400.42	7.8*	183.19	298.95	337.59	362.18	382.33	402.48	421.16	439.54	464.51	500.76	635.98
400.99	2.02*	152.84	299.55	346.36	371.24	391.16	409.16	424.45	438.92	457.73	484.39	599.54
401.04	1.63	258.91	342.78	363.85	378.64	392.02	403.56	412.59	426.41	438.26	455.65	515.54
407.34	3.9*	162.96	300.15	343.89	372.9	395.58	413.56	431.08	450.05	469.37	497.44	595.34
407.46	2.93	220.32	314.19	344.3	368.36	385.39	403.4	425.48	449.96	471.44	502.89	589.64
409.65	2.66*	149.78	315.67	356.93	382.11	398.72	414.26	430.81	447.26	464.84	494.61	594.11
411.81	3.19*	173.85	315.81	355.87	382	401.35	417.68	432.41	449.51	469.16	497.48	605.63
421.16	2.81*	187.59	326.3	362.07	385.5	407.12	424.45	443.01	459.63	481.73	505.94	634.18
421.74	2.77*	192.68	336.15	373.65	395.33	413.02	427.06	441.69	454.44	471.01	494.62	570.98
424.11	3.65	224.77	333.53	364.4	383.24	403.91	419.4	439.74	458.89	483	518.93	594.88
426.12	3.34*	183.38	339.3	371.61	393.54	412.53	428.24	442.83	462.76	483.35	515.21	618.77
427.54	4.43*	198.88	328.67	364.13	392.39	411.34	431.63	449.19	467.48	487.96	518.17	633.28
433.29	3.59*	204.37	345.26	380.6	401.65	418.79	434.99	449.1	466.1	486.56	523.79	628.13
434.38	2.63*	188.21	334.82	370.79	396.93	419.31	438.86	455.42	475.13	501.91	526.3	648.28
437.23	3.03*	191.05	342.53	378.26	403.58	424.66	441.8	459.01	473.95	496.25	527.41	616.47
442.39	3.51	250.4	349.94	382.51	408.77	426.25	443.87	462.35	479.27	501.23	527.64	627.97
462.65	2.42*	193.39	373.09	407.46	433.4	452.04	466.28	481.15	499.76	518.23	546.32	684.15
477.27	2.85	270.53	400.39	432.52	451.07	468.31	481.81	496.42	510.26	526.58	548.15	670.91
491.62	1.97	284.05	420.75	452.84	470.07	483.75	495.44	506.22	518.88	534.03	549.89	624.32
491.85	3.89	277.32	404.8	435.05	457.15	475.91	492.7	508.54	526.45	545.82	572.57	685.99
492.78	2.5	292.07	426.38	450.49	466.75	480.74	494.36	508.38	521.49	535.65	555.64	636.41
501.61	1.58	294.14	428.66	458.93	477.05	490.82	504.02	516.2	530.37	547.28	570.69	660.67
503.17	2.53	269.1	428.57	458.99	478.35	492.39	507.16	521.15	534.26	550.17	572.58	649.6
504.75	2.33	293.34	425.05	454.79	474.83	493.06	508.66	522.82	536.37	551.51	575.69	650.8
505.41	2.69	305.02	436.54	459.63	475.85	491.7	505.45	518.06	533.75	550.18	575.07	665.12
508.60	2.02	257.4	436.49	465.37	485.11	500.75	512.49	524.86	537.97	553.37	573.86	652.7
511.72	1.67	283.82	441.21	467.56	484.24	500.83	513.54	525.32	539.79	556.74	580.32	678.45
512.36	2.79	237.55	430.4	460.45	483.36	501.75	518.54	531.16	547.12	564.78	586.85	694.78
512.60	1.24	283.65	435.42	468.04	488.96	503.51	518.04	531.36	544.17	559.84	579.22	647.03
514.67	7.03	281.3	428.78	466.16	484.02	499.37	515.81	531.41	548.28	570.28	599.71	708.92
515.51	2.79	305.98	437.54	466.21	487.24	503.38	517.69	532.91	549.25	565.47	587.34	672.02
517.44	2.3	260.66	446.55	475.24	494.08	507.92	520.07	532.79	545.74	561.52	583.45	672.68
521.27	2.03	296.6	461.58	485.39	499.89	511.7	522.03	533.6	544.67	559.6	581.12	640.71
522.03	2.31	288.29	450.34	476.13	495.41	509.69	523.07	536.03	550.5	567.34	590.99	685.49
522.99	2.48	303.41	447.82	477.19	499.03	513.81	527.55	541.21	555.61	569.85	591.07	688.79
523.11	2.67	288.95	443.71	471.78	495.36	513.51	529.59	543.2	556.84	571.71	592.7	671.52
525.66	2.5	288.39	445.59	475.44	496.44	513.3	529.01	544.99	558.96	576.56	598.29	675.2
530.96	3.34	277.82	459.16	486.32	504.37	518.64	530.33	543.5	557.22	575.16	602.92	685.15
531.15	2.17	311.03	461.64	487.58	503.2	516.43	529.79	542.08	557.72	574.02	601.26	680.61
534.41	3.32	369.15	471.85	496.64	510.36	521.73	531.31	545.97	557.95	568.25	582.47	623.23
537.72	2.28	332.12	474.66	499.51	515.5	527.35	538.74	550.82	562.16	577.75	598.14	682.6
538.88	2.68	296.35	451.82	485.63	507.75	527.54	542.65	557.38	572.17	593.72	619.7	725.04
544.63	2.94	299.35	467.72	496.51	518.11	536.2	548.69	562.98	577.16	592.08	613.33	705.32
552.78	3.73	339.55	487.93	514.17	531	545.01	556.64	569.29	582.76	596.66	616.93	712.96
557.19	2.58	265.81	458.05	497.84	523.48	542.78	561.19	579.1	596.79	617.95	646	746.78
561.82	2.79	229.5	437.83	481.47	510.77	540.16	567.57	594.4	617.13	641.98	678.49	815.19
582.47	4.18	367.86	510.76	536.91	553.67	566.83	580.7	595.57	610.22	627.42	653.95	747.57

Table 8 **Science, age 9 and 13, mean proficiency scores and percentile by population**

MEAN	SE	C 1	C 10	C 20	C 30	C 40	MED	C 60	C 70	C 80	C 90	C 100
374.41	4.46*	140.22	269.09	309.39	332.55	354.34	374.84	394.09	415.79	441.05	474.02	572.64
382.05	4.07*	119.03	249.51	301.91	338.6	364.42	389.76	413.1	436.63	462.23	493.67	605.63
383.63	2.63*	130.02	285.61	322.64	348.96	366.65	381.73	400.1	423.11	444.99	475.82	587.61
385.25	2.56*	169.31	291.91	327.2	353.9	370.6	387.54	404.5	424.69	447.07	473.1	568.35
412.25	4.34	169.13	301.78	337.98	365.65	387.62	410.12	432.34	458.19	487.5	522.02	675.29
415.57	2.32*	134.5	283.85	348.39	378.19	407.24	428.44	446.81	463.66	486.12	516.15	616.07
417.15	4.25*	156.16	302.1	344.6	373.33	398.52	423.03	444.48	466.87	491.4	525.96	643.69
418.08	3.74*	132.6	305.04	343.88	375.01	396.09	416.61	440.74	464.69	492.68	526.54	625.6
420.43	3.62*	160.61	308.9	356.01	379.73	405.72	423.13	445.36	465.64	489.69	517.03	605.66
421.35	3.25*	154.75	302.13	351.81	381.26	407.34	431.77	450.7	469.71	492.94	523.77	630.09
421.89	6.05*	161.45	328.85	360.49	382.74	400.44	419.75	437.82	456.49	483.08	517.31	635.89
424.57	3*	149.72	331.31	362.64	389.48	409.28	429.94	446.53	465.67	485.73	513.73	640.56
425.91	5.*69	152	295.71	348.07	384.03	410.09	433.88	454.51	479.24	502.48	532.5	641.36
426.43	2.8*7	156.35	333.34	369.66	394.25	412.85	429.47	447.85	464.79	486.07	511.89	598.19
426.62	3.97*	163.91	308.2	354.54	386.53	409	431.82	453.55	472.8	498.27	533.26	625.27
435.63	5.45*	141.93	295.25	354.71	393.63	421.04	444.18	468.5	489.75	517.14	550.88	683.02
445.52	3.4	203.01	326.05	367.69	396.46	422.52	444.8	467.79	486.96	513.79	550.91	668.8
446.32	3.5*	179.23	343.37	384.75	414.54	433.93	454.16	471.67	488.2	508.12	535.62	633.81
447.77	3.21*	160.14	331.5	375.91	405.79	427.26	449.58	470.95	496.56	520.78	561.39	664.22
451.82	4.96*	175.33	346.21	383.84	405.5	429.99	452.2	474.51	495.44	519.71	552.97	651.2
453.04	2.79*	207.95	360.33	395.49	416.28	434.59	451.87	470.23	490.14	513.77	546.85	645
467.93	3.89	164.65	351.34	394.72	427.97	454.39	475.34	495.74	517.2	538.97	566.39	664.19
496.29	2.64	274.93	407.55	438.43	461.01	477.17	494.93	514.48	534.28	556.81	584.4	685.65
507.53	4.04	227.89	402.27	444.48	469.11	490.64	510.4	528.69	551.62	577.12	606.24	759.49
511.31	3.01	227.36	402.75	446.06	472.41	494.09	512.67	532.69	555.79	578.95	610.72	723.8
511.54	1.73	232.71	415.63	453.62	475.82	494.96	515.78	532.16	549.74	570.49	600.81	709.82
520.45	4	292.96	451.86	470.2	487.14	500.62	514.82	530.5	545.79	566.94	597.45	663.03
524.8	5.21	182.23	423.9	466.56	489.91	509.14	527.06	544.06	561.81	584.13	617.59	727.17
524.87	2.31	268.86	429.56	466.29	489.56	508.64	525.39	544.35	562.52	583.41	618.81	706.96
524.91	1.77	240.86	430.43	466.4	491.41	511.46	529.24	546.89	564.25	585.48	613.99	736.33
526.74	3.37	296.63	439.23	470.2	492.22	510.01	528.05	544.28	560.67	581.37	610.49	682.67
529.33	2.77	247.75	444.3	473.12	495.02	513.01	528.58	545.25	562.67	582.22	613.05	712.85
530.36	3.32	275.43	443.49	478.33	499.91	516	530.39	544.92	560.48	581.71	611.21	712.81
531.12	5.6	248.69	426.94	466.17	493.33	514.52	533.35	551.21	569.98	595.36	627.5	747.72
535.17	3.36	252.46	430.19	472.87	497.86	520.6	538.27	558.23	575.81	599.29	631.11	749.64
536.55	3.15	260.02	434.92	472.69	504.21	522.72	539.49	558.41	576.14	599.22	631.57	762.48
537.99	2.99	289.44	430.79	474.59	500.24	520.98	540.1	558.85	580.02	602.23	630.95	735.12
538.13	2.27	300.91	443.63	475.63	501.74	522.25	540.03	558.68	578.4	599.5	628.81	735.93
538.22	6	226.85	430.95	472.43	495.8	519.61	541.22	559.49	579.85	603.79	634.08	719.43
540.66	3.37	274.45	439.41	475.47	502.35	523.83	542.3	562.99	583.1	601.99	634.52	745.32
542.13	2.58	282.22	455.54	486.13	508.33	525.46	542.32	558.08	575	597.36	630.15	735.36
543.85	2.95	276.07	451.33	484.28	508.7	527.63	544.74	563	582.1	603.4	632.1	729.98
544.16	2.57	254.09	453.16	483.05	506.37	525.41	546.22	562.02	581.34	605.04	633.9	728.35
544.59	2.65	264.51	457.77	489.06	510.36	529.63	546.95	561.33	580.4	602.53	630.69	730.26
548.17	4.2	274.19	457.57	490.27	514.61	535.66	553.57	570.09	588.6	605.15	632.38	700.41
552.16	2.57	301.09	470.3	498.57	517.45	535.25	550.76	566.55	584.71	608.98	633.77	724.21
557.58	2.26	288.7	473.88	507.4	528.55	543.89	561.11	575.16	590.22	607.93	633.74	730.51
563.13	2.94	297.35	461.29	496.45	520.99	540.99	564.78	586.57	606.5	628.55	664.21	779.82
564.36	2.14	242.1	479.9	509.88	531.42	549.61	566.29	583.61	599.31	619.45	646.86	816.52
573.87	4.26	309.41	489.19	522.52	543.83	560.47	578.03	593.31	609.63	632.06	663.02	785.75
575.49	2.23	251.2	455.41	500.73	537.77	563.82	585.28	606.42	628.49	650.36	682.75	786.11
583.82	2.72	266.76	487.23	524.28	550.35	571.04	589.38	605.71	622.55	645.11	675.76	771.89

Table 9 Mathematics, equated slope parameters for common items

	9-year-olds	13-year-olds	
Item17	0.76897	0.73187	Item20
Item18	0.70382	0.83575	Item21
Item19	0.77018	0.66709	Item22
Item20	0.86174	0.59672	Item23
Item21	0.71834	0.56451	Item24
Item22	0.90387	0.67999	Item25
Item23	0.87985	1.03708	Item26
Item24	1.1638	0.99587	Item27
Item25	0.79035	0.82189	Item28
Item26	0.95702	0.99254	Item29
Item27	0.88049	1.11276	Item30
Item28	1.15055	1.24063	Item31
Item29	0.95931	0.87039	Item32
Item30	0.95709	1.07588	Item33

Table 10 Mathematics, equated treshold parameters for common items

	9-year-olds	13-year-olds	
Item17	-1.36162	-1.44386	Item20
Item18	-1.81124	-1.66592	Item21
Item19	-0.47919	-0.14352	Item22
Item20	-0.24729	-0.41038	Item23
Item21	-0.76059	-0.56463	Item24
Item22	-1.33062	-1.45374	Item25
Item23	-0.98547	-0.58454	Item26
Item24	-0.16388	-0.38028	Item27
Item25	-0.45516	-0.33349	Item28
Item26	-0.84136	-0.74667	Item29
Item27	0.43431	-0.62607	Item30
Item28	-0.05627	0.23263	Item31
Item29	-0.4826	-0.60479	Item32
Item30	-0.23453	-0.53731	Item33

Table 11 Mathematics, equated guessing parameters for common items

	9-year-olds	13-year-olds	
Item17	0.23484	0.20564	Item20
Item18	0.25686	0.18289	Item21
Item19	0.27117	0.3104	Item22
Item20	0.28042	0.18696	Item23
Item21	0.27469	0.20861	Item24
Item22	0.16562	0.18137	Item25
Item23	0.19152	0.30164	Item26
Item24	0.30381	0.17997	Item27
Item25	0.17341	0.17542	Item28
Item26	0	0	Item29
Item27	0.32057	0.37623	Item30
Item28	0.27199	0.2069	Item31
Item29	0.1609	0.24596	Item32
Item30	0.23691	0.24811	Item33

Table 12 **Science, equated slope parameters for common items**

	9-year-olds	13-year-olds	
Item16	0.16192	0.3003	Item16
Item17	0.68524	0.74228	Item17
Item18	0.75434	0.76781	Item18
Item19	0.72529	0.66119	Item19
Item20	0.55497	1.03132	Item20
Item21	0.80861	0.71951	Item21
Item22	0.53808	0.74359	Item22
Item23	0.60036	0.6129	Item23
Item24	0.48032	0.64799	Item24
Item25	0.58275	0.95714	Item25
Item26	0.76455	0.37457	Item26
Item27	0.29778	0.7748	Item27
Item28	0.74936	0.86323	Item28
Item29	0.94014	1.32178	Item29

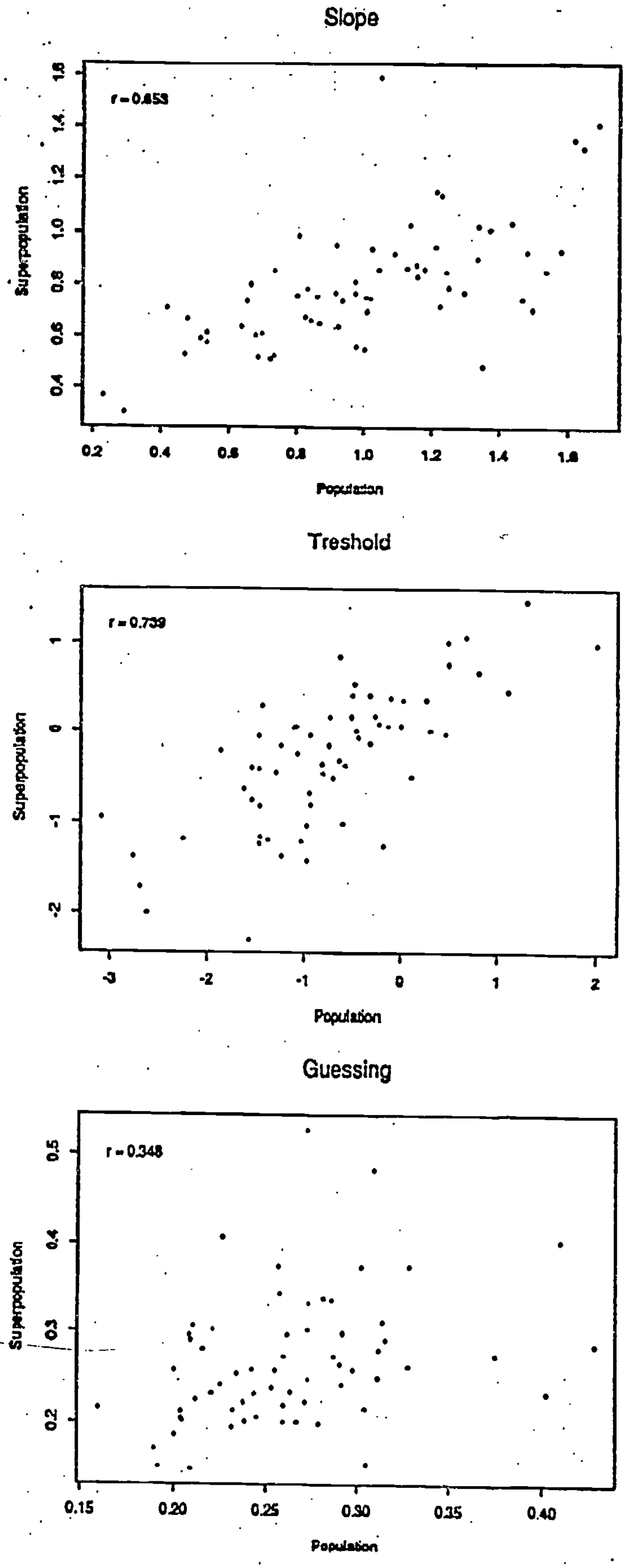
Table 13 **Science, equated treshold parameters for common items**

	9-year-olds	13-year-olds	
Item16	-2.64404	-1.36454	Item16
Item17	-0.79376	-0.81066	Item17
Item18	-1.1249	-1.00337	Item18
Item19	-1.29098	-1.4272	Item19
Item20	-1.05326	-0.21995	Item20
Item21	-0.13286	-0.38295	Item21
Item22	-1.35272	-1.19196	Item22
Item23	-1.48899	-1.17802	Item23
Item24	-2.36079	-2.00635	Item24
Item25	-0.61151	-0.47411	Item25
Item26	-1.00494	0.17538	Item26
Item27	-0.78997	-0.14441	Item27
Item28	-0.63007	-0.47186	Item28
Item29	-0.25141	-0.5232	Item29

Table 14 **Science, equated guessing parameters for common items**

	9-year-olds	13-year-olds	
Item16	0.29306	0.34093	Item16
Item17	0.29953	0.34522	Item17
Item18	0.30201	0.26712	Item18
Item19	0.21697	0.26014	Item19
Item20	0.36796	0.48258	Item20
Item21	0.3591	0.37624	Item21
Item22	0.22984	0.31043	Item22
Item23	0.17459	0.23425	Item23
Item24	0.31988	0.27574	Item24
Item25	0.27565	0.40488	Item25
Item26	0.18804	0.30189	Item26
Item27	0.27049	0.41344	Item27
Item28	0.22244	0.255	Item28
Item29	0.27053	0.24219	Item29

Figure 1 **Reference population item parameter estimates compared to population x item parameter estimates**



Slope

Treshold

Guessing

Figure 2 Reference population item parameter estimates compared to population y item parameter estimates

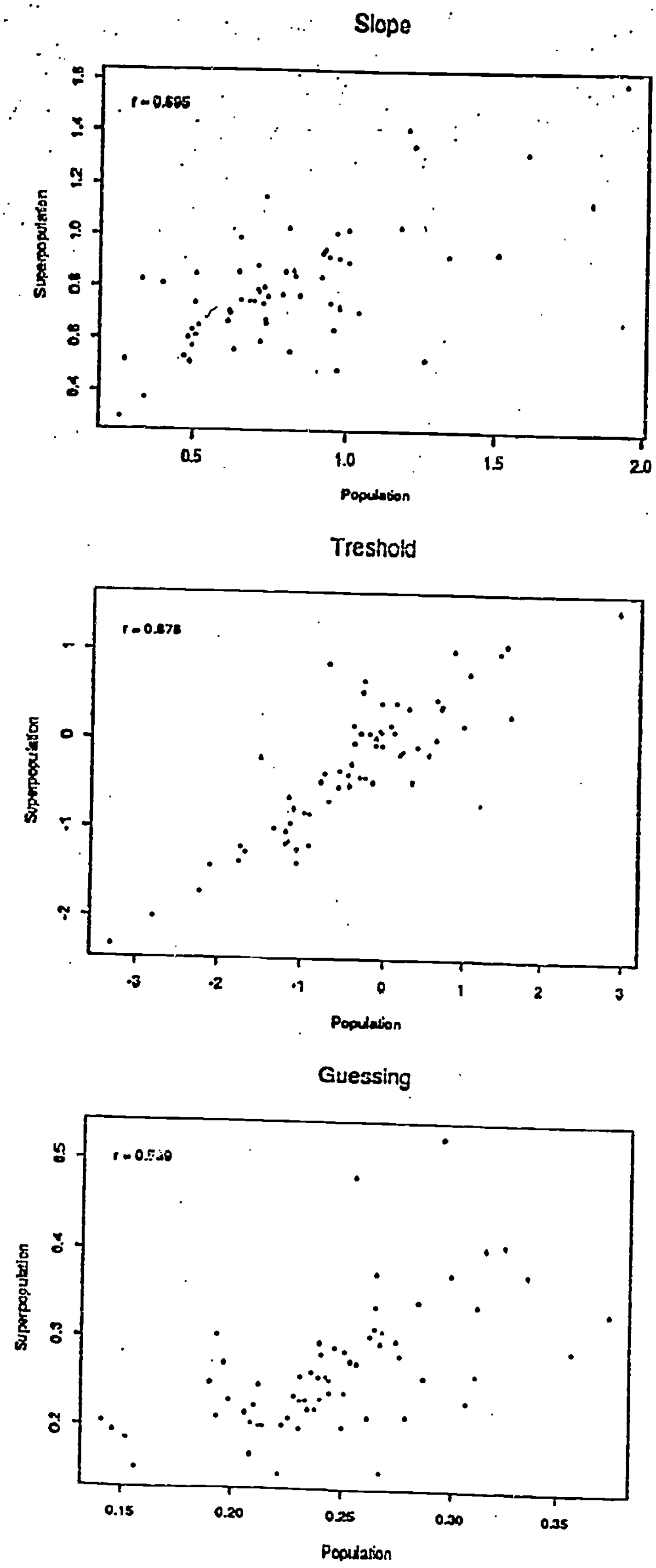
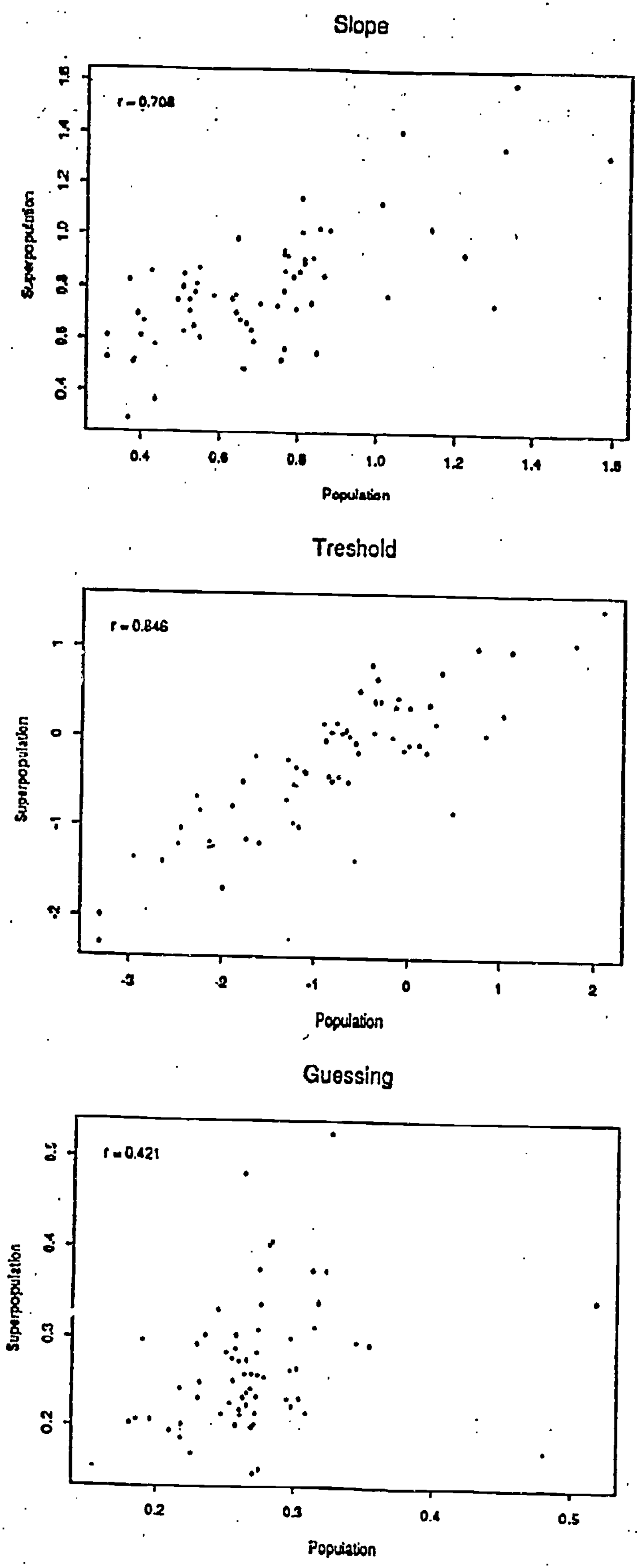


Figure 3 Reference population item parameter estimates compared to population z item parameter estimates

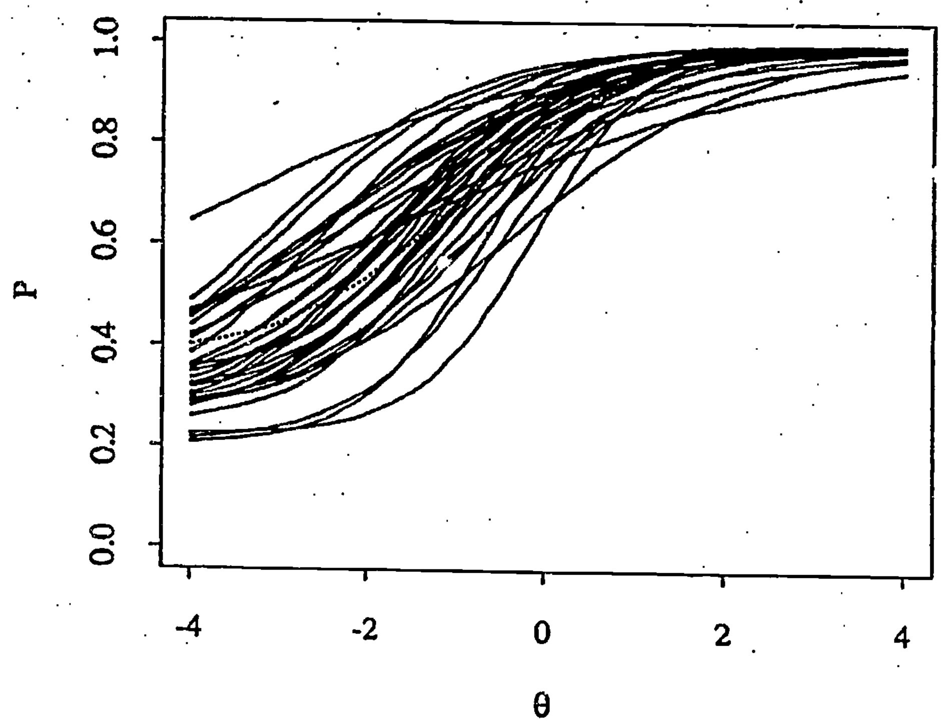


Slope

Treshold

Guessing

Figure 4 Science, age 13, items 1 and 2, superimposed ICCs estimated from each individual population (solid lines) and the reference population (dashed line)



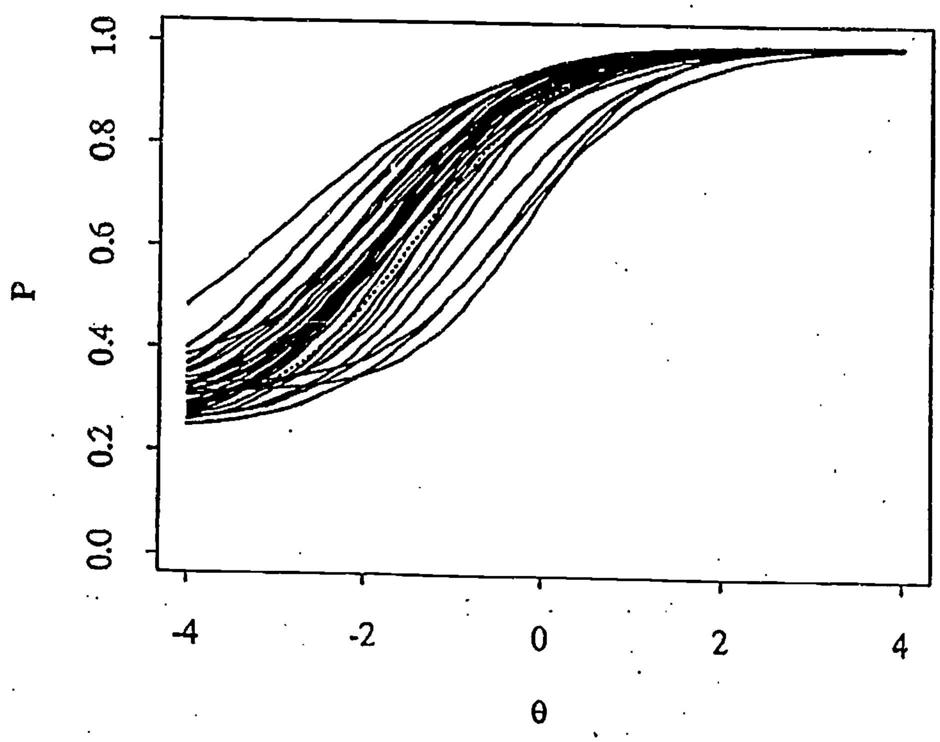


Figure 5 Science, age 13, items 3 and 4, superimposed ICCs estimated from each individual population (solid lines) and the reference population (dashed line)

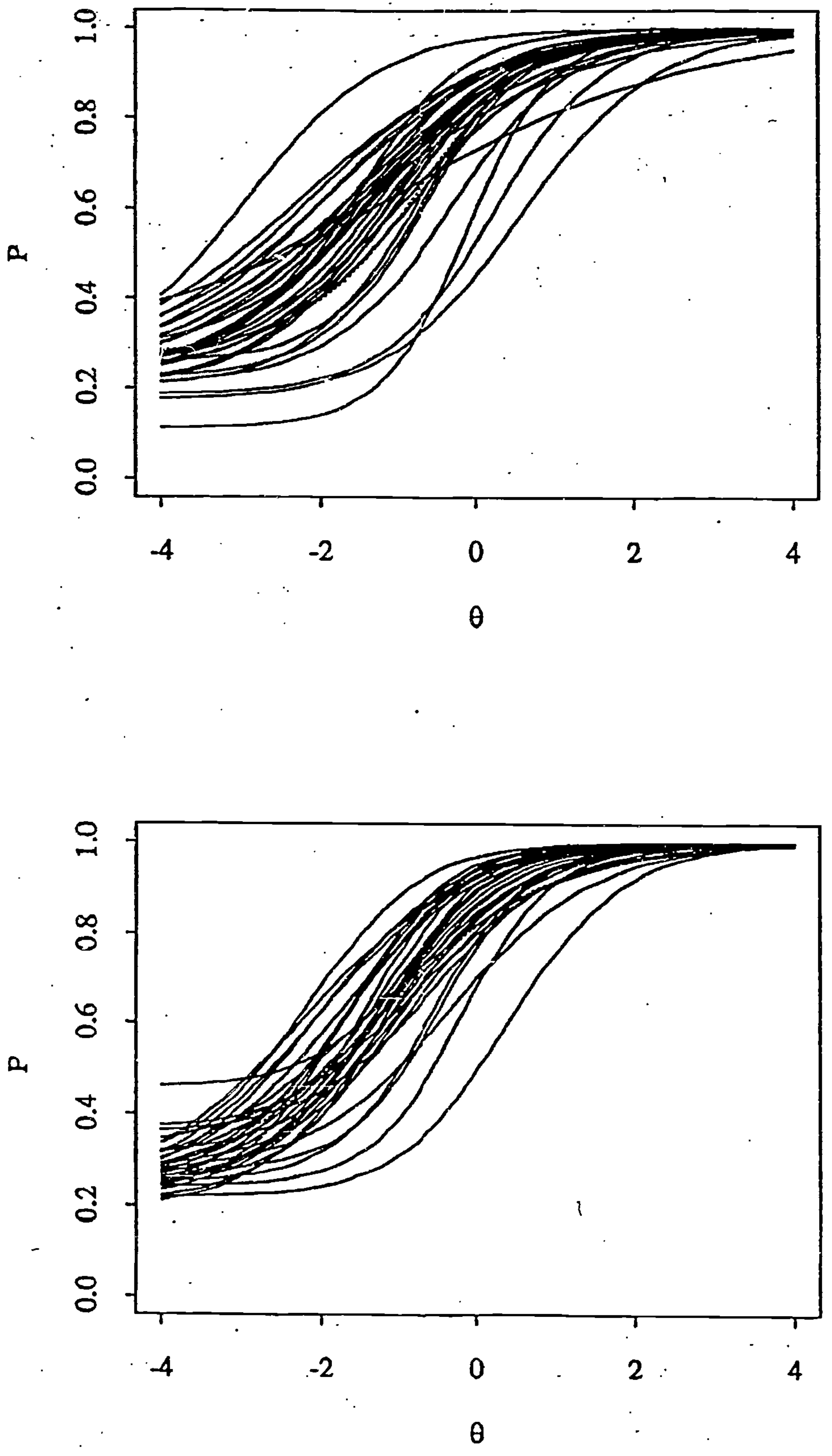
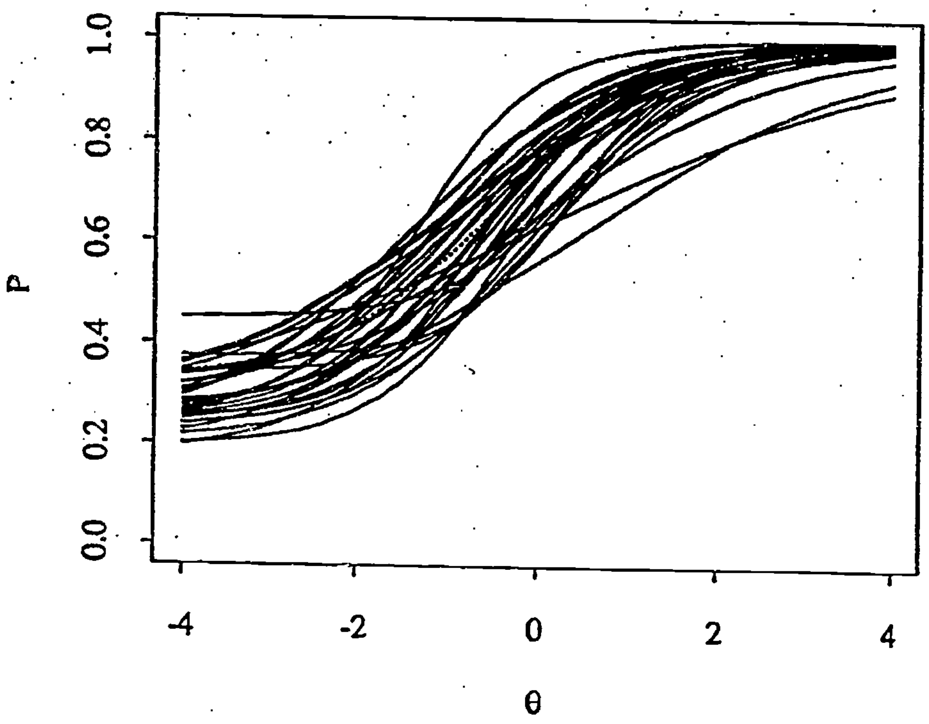


Figure 6 Science, age 13, items 5 and 6, superimposed ICCs estimated from each individual population (solid lines) and the reference population (dashed line)



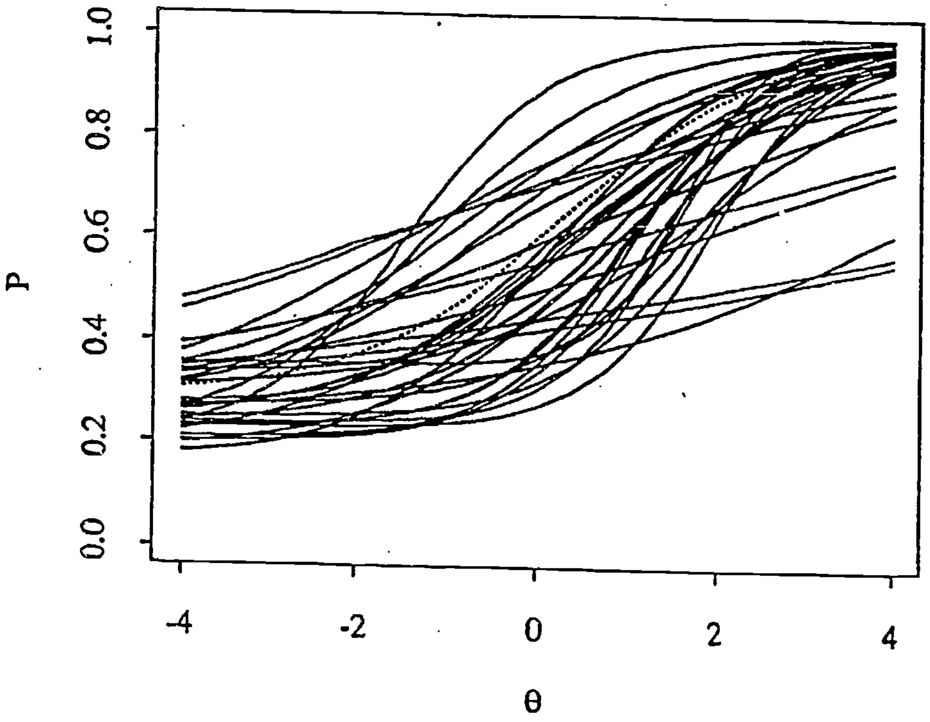


Figure 7 **Mathematics, ages 9 and 13, superimposed regression lines**

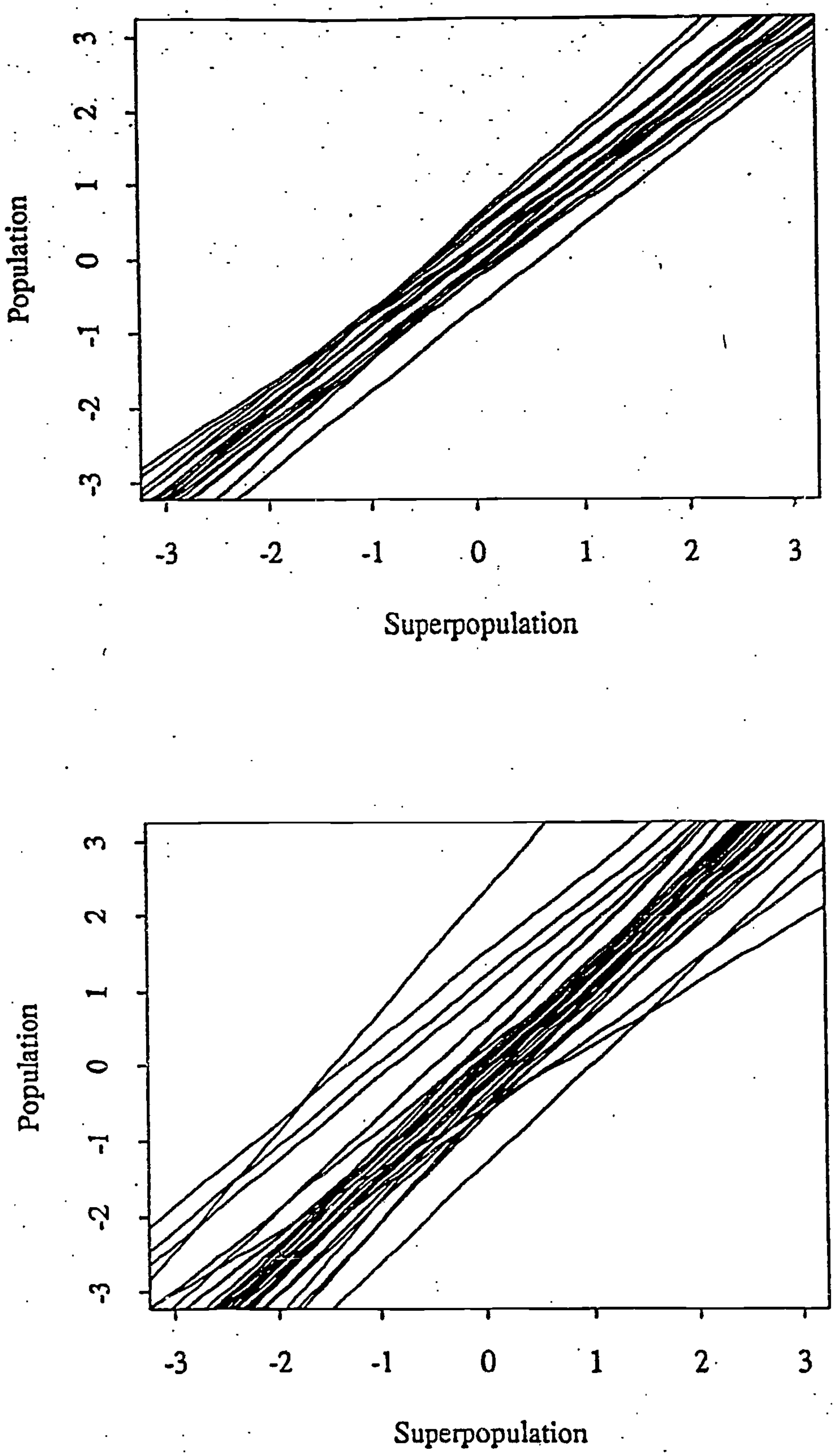
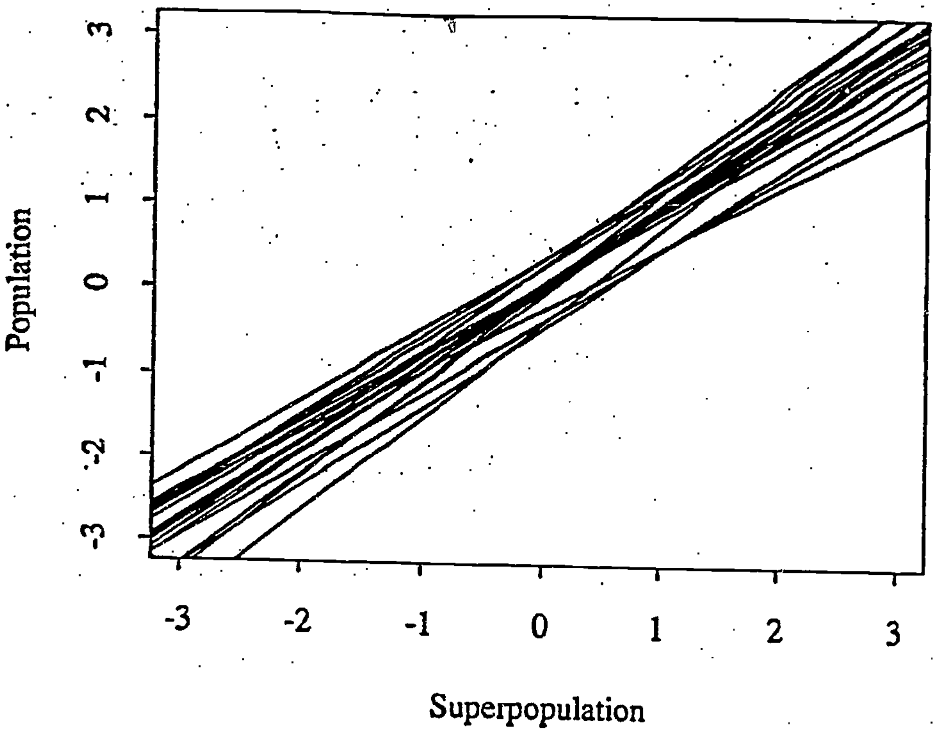


Figure 8 Science, ages 9 and 13, superimposed regression lines



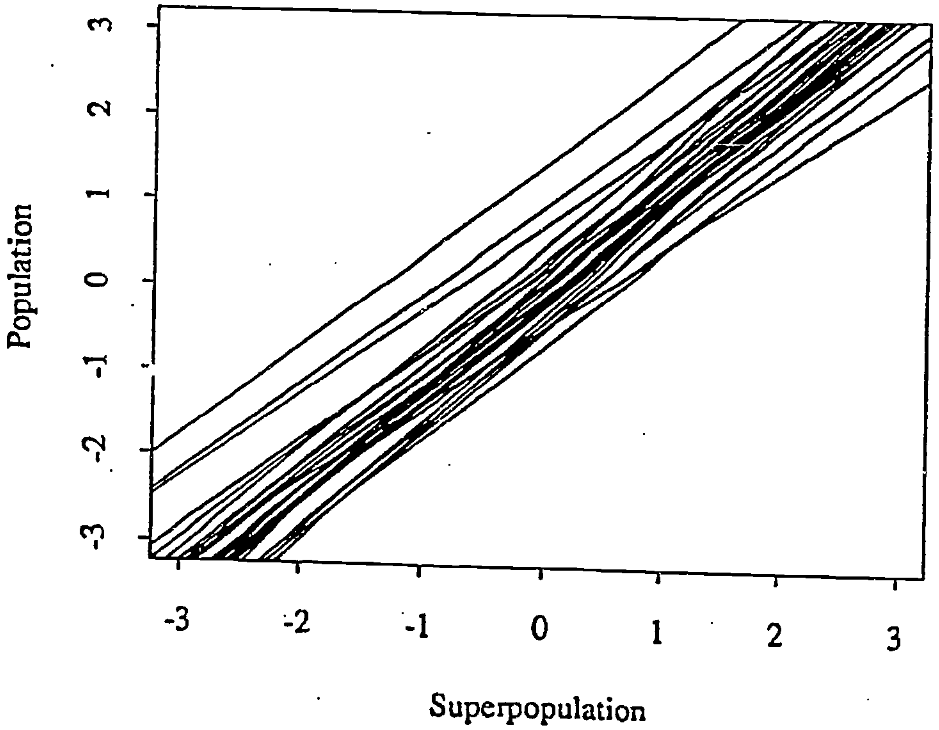


Figure 9 Mathematics, ages 9 and 13, superimposed regression lines for equated population item parameter estimates

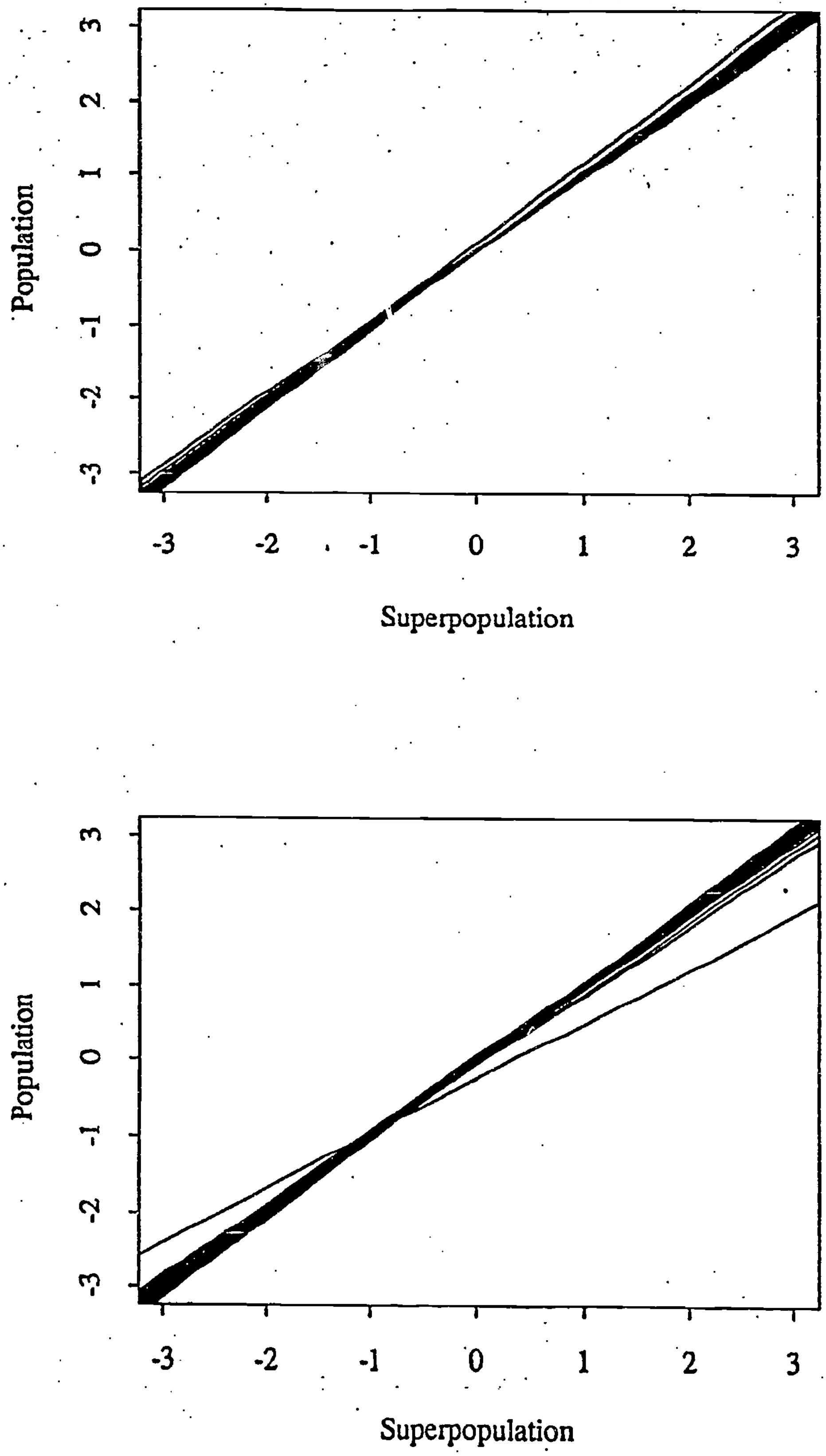
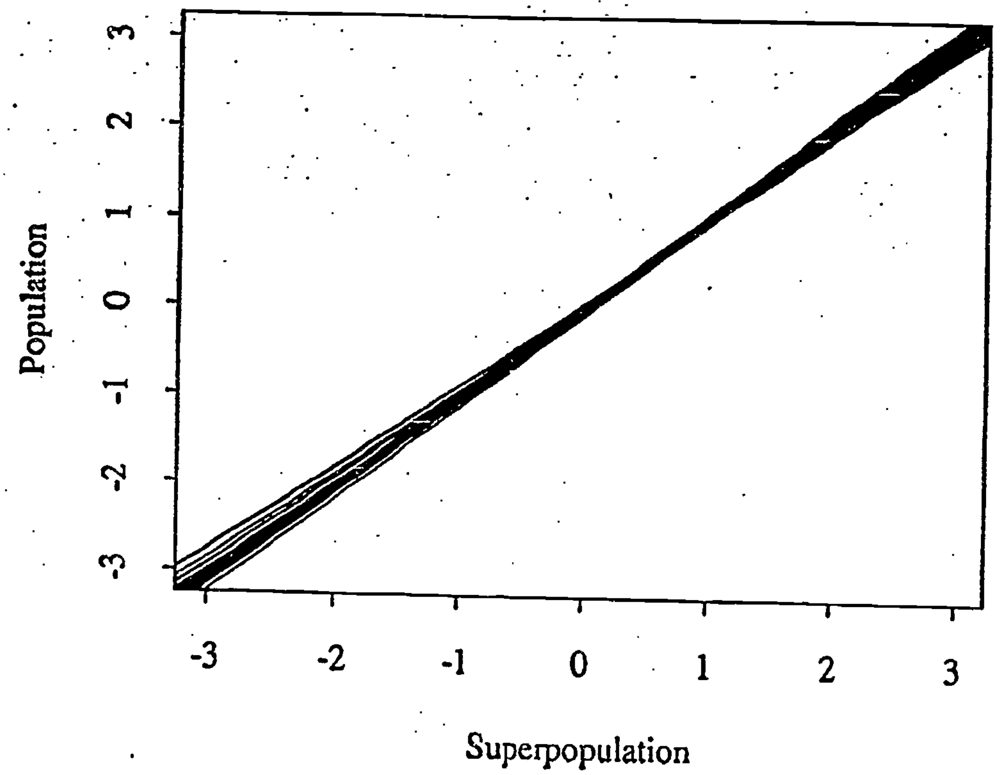


Figure 10 Science, ages 9 and 13, superimposed regression lines for equated population item parameter estimates



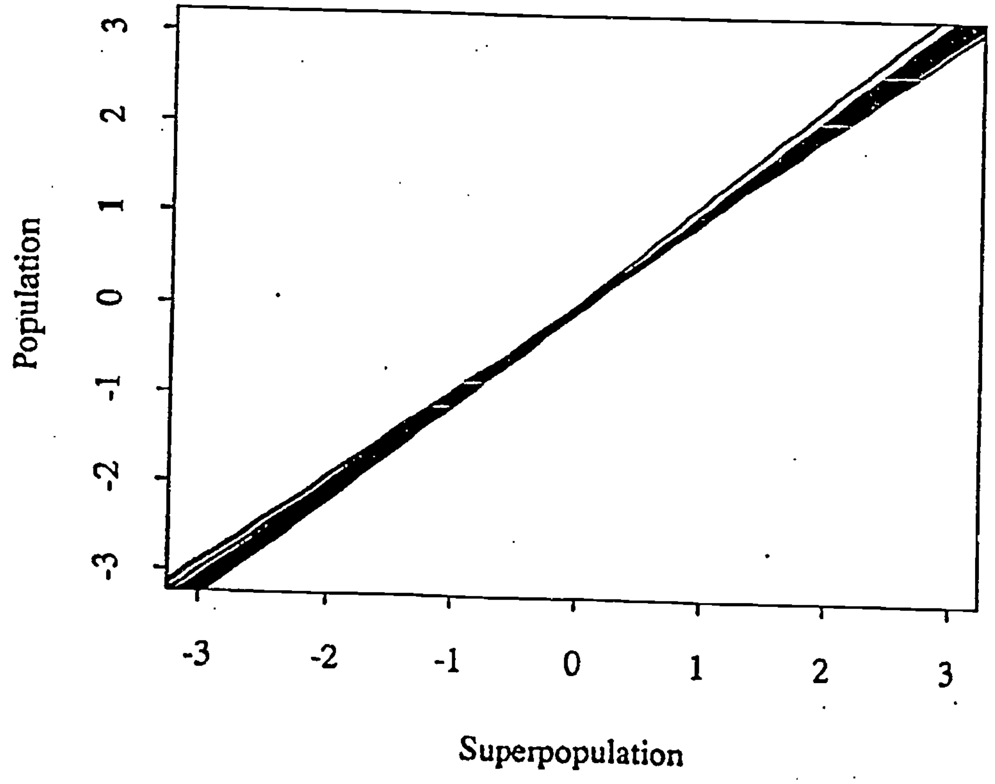


Figure 11 Mathematics, equated slope parameters for the common items.

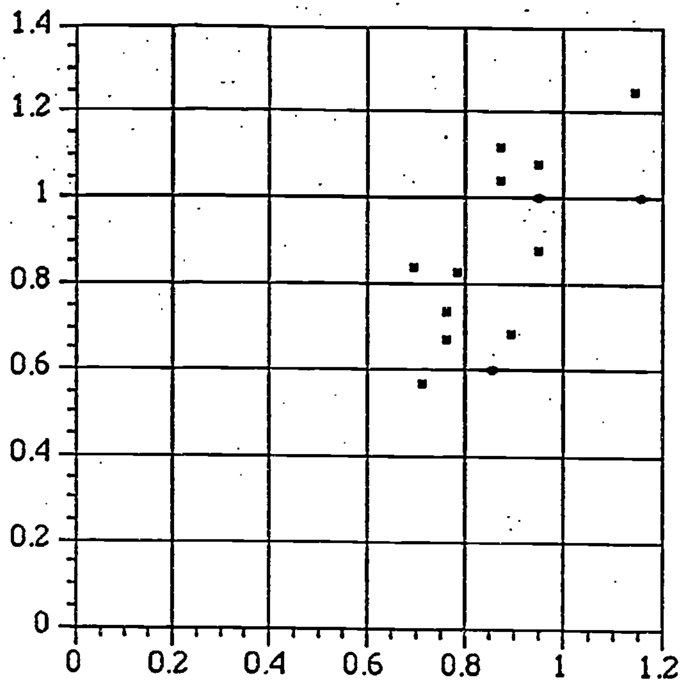


Figure 12 Mathematics, equated threshold parameters for the common items.

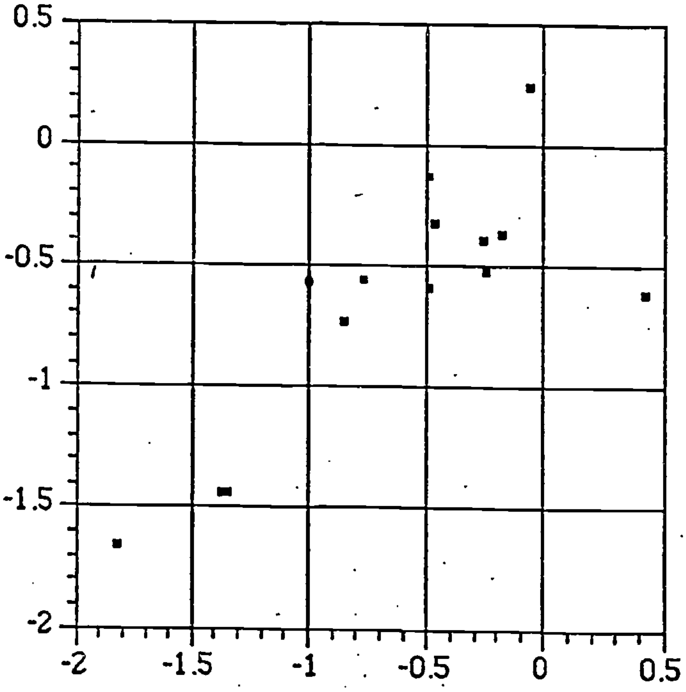


Figure 13 Mathematics, equated guessing parameters for the common items.

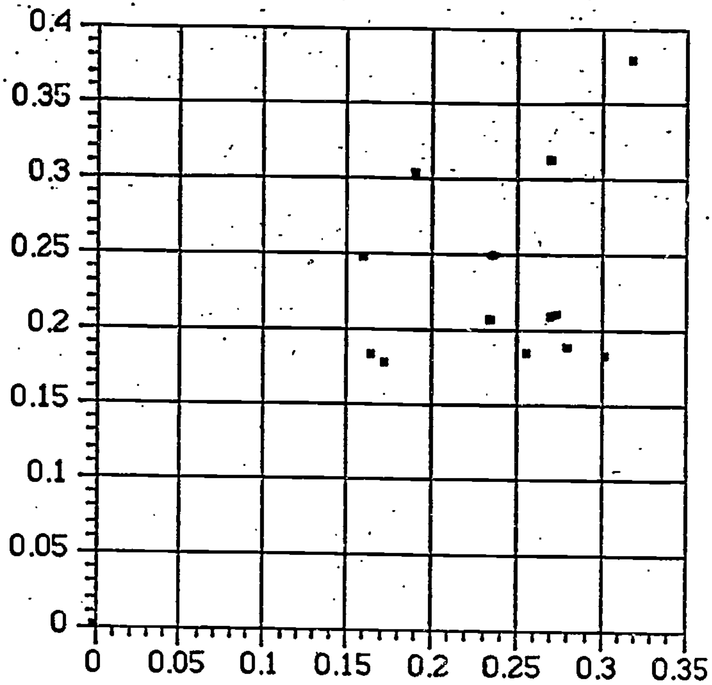


Figure 14 Science, equated slope parameters for the common items.

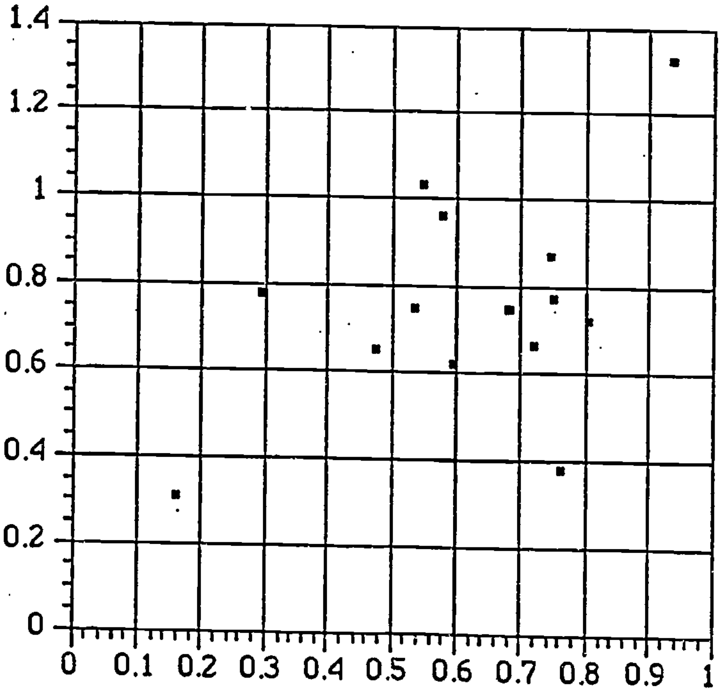


Figure 15 Science, equated threshold parameters for the common items.

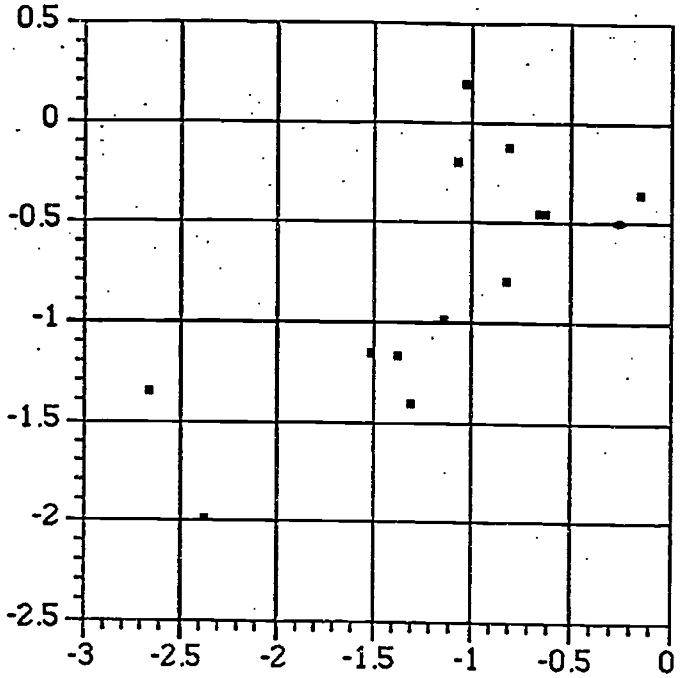
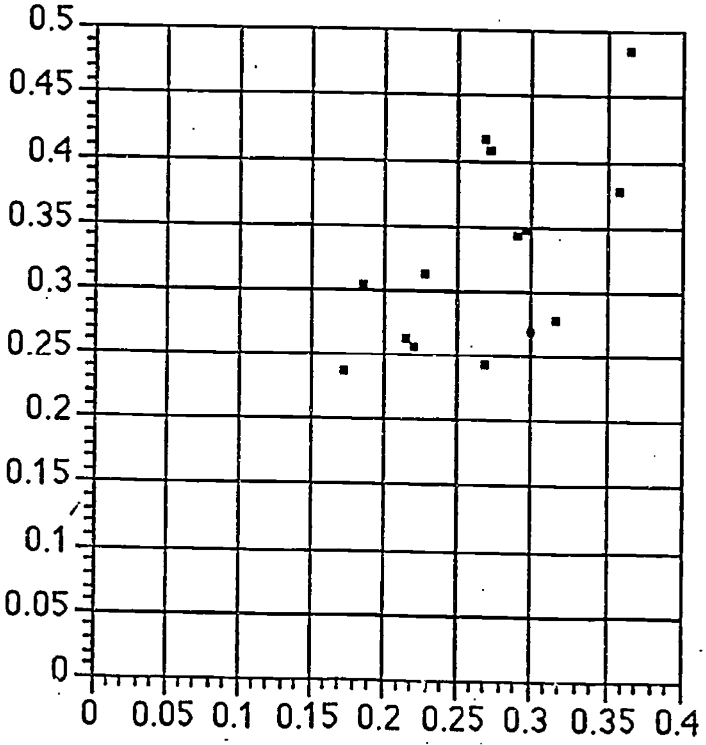


Figure 16 Science, equated guessing parameters for the common items.


0.5
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
0
0.05
0.1
0.15
0.2
0.25
0.3
0.35
0.4