Journal of Educational Measurement
Summer 2008, Vol. 45, No. 2, pp. 139–157

Comparing the Difficulty of Examination Subjects with Item Response Theory

Oksana B. Korobko and Cees A. W. Glas
University of Twente, the Netherlands
Roel J. Bosker
University of Groningen, the Netherlands
Johan W. Luyten
University of Twente, the Netherlands

Methods are presented for comparing grades obtained in a situation where students can choose between different subjects. It must be expected that the comparison between the grades is complicated by the interaction between the students' pattern and level of proficiency on one hand, and the choice of the subjects on the other hand. Three methods based on item response theory (IRT) for the estimation of proficiency measures that are comparable over students and subjects are discussed: a method based on a model with a unidimensional representation of proficiency, a method based on a model with a multidimensional representation of proficiency, and a method based on a multidimensional representation of proficiency where the stochastic nature of the choice of examination subjects is explicitly modeled. The methods are compared using the data from the Central Examinations in Secondary Education in the Netherlands. The results show that the unidimensional IRT model produces unrealistic results, which do not appear when using the two multidimensional IRT models. Further, it is shown that both the multidimensional models produce acceptable model fit. However, the model that explicitly takes the choice process into account produces the best model fit.

Introduction

The problem of grade adjustment for the comparison of students and schools has a long history (see, for instance, Linn, 1966). Johnson (1997, 2003) notes that combining student grades through simple averaging schemes to obtain grade point averages (GPAs) results in systematic bias against students enrolled in more rigorous curricula. The practice has important consequences for the course selection by the students, and it may be one of the major causes of grade inflation. Caulkins, Larkey and Wei (1996) note that the use of GPA is based on the incorrect assumption that all course grades mean essentially the same thing. There is, however, substantial variation among majors, courses, and instructors in the rigor with which grades are assigned. A lower GPA may not necessarily mean that the student performs less well than students who have higher GPAs; the students with lower GPAs may simply be taking courses and studying in fields with more stringent grading standards.

The appropriateness of GPAs is also a point of debate in school effectiveness research and in the trend towards public reporting of school results. School results are generally corrected for differences between the students at school entry (Fitz-Gibbon, 1994; Willms, 1992), but the comparability of the actual outcome measures, such as examination results, has received less attention, with the exceptions of Kelly (1976), Newton (1997), and Smits, Mellenbergh and Vorst (2002). In many countries (such as the Netherlands, where the data used here emanate) a student's examination result has a direct consequence for the admittance to university. Therefore, students generally choose the examination subjects in which they feel competent. The focal problem addressed by Kelly, Newton, and Smits, and Mellenbergh and Vorst is whether the fact that students generally choose subjects that fit their proficiency distorts the comparison of average examination results between schools. Parents, local authorities and politicians, however, may interpret these differences in GPAs as absolute objectivity, ignoring the influence of the differences in the difficulty of the subjects and the students' choice behavior.

Most more recent methods for adjusting GPA are based on item response theory (IRT). The objective of these methods is to account for the relative difficulty of the courses or examinations and the differences in the proficiency levels of the students (Johnson, 1997, 2003; Young, 1990, 1991). In the present article, this approach is expanded in two directions. First, it is assumed that the courses or examinations load on more than one dimension. (In the sequel, we will use the term examinations as a generic name that also includes assessments of courses etc.). Using a real-data example it is shown that a multidimensional representation of proficiency leads to more plausible results and better model fit. Second, it is argued that the free choice of examinations may lead to a violation of the ignorability principle (Rubin, 1976) and, as a consequence, may lead to biased estimates of the difficulties of the examination subjects. It is shown that this bias can, to a certain extent, be accounted for by introducing a stochastic model for the choice variables.

This article is organized as follows. First, three IRT models will be described: a unidimensional and a multidimensional model for the grades only, and a multidimensional model pertaining to the grades and the choice variables simultaneously. As an example, an analysis of data collected by Dutch Inspection of Education will be presented. Then, a method for the evaluation of model fit will be described and the fit of the three models will be compared. Finally, the last section presents a discussion and some conclusions.

Method

Grade Point Average Adjustment

One might view the problem of comparing the difficulty of examinations as an item scaling problem with incomplete data where an item score is the (discrete, polytomous) score on an examination subject. We define a choice variable as

$$d_{ni} = \begin{cases} 1 & \text{if student } n \text{ did chose examination subject } i \\ 0 & \text{if student } n \text{ did not chose examination subject } i, \end{cases} \tag{1}$$

for students $n = 1, \ldots, N$ and examination subjects $i = 1, \ldots, K$. An important aspect of the problem discussed in this article is that the design (that is, the values of the choice variables d_{ni}) is not fixed in advance, but it is student driven and therefore stochastic. The consequences of the stochastic nature of the design will be returned to below.

The objective is to compute adjusted GPAs in such a way that they are comparable. This is done by estimating the GPA for a situation where all students take all examinations. Since they do not actually take all examinations, we impute expected grades for the missing observations, that is

$$GPA = \frac{1}{K}\sum_{i=1}^{K}(d_{ni}X_{ni} + (1 - d_{ni})E(X_{ni})), \tag{2}$$

where X_{ni} is the observed grade if $d_{ni} = 1$ and an arbitrary value if $d_{ni} = 0$, and $E(X_{ni})$ is the expectation under a model used to describe the students' proficiency.

Item Response Theory

The expectations $E(X_{ni})$ in equation (2) will be computed using IRT models for the proficiency of the students and the difficulty of the examination subjects. Three models will be discussed. In the first model, it will be assumed that the grades on all subjects can be modeled using a unidimensional representation of proficiency. In the second model, this assumption is broadened to the assumption that the subjects relate to more than one proficiency dimension. The third model is motivated by the expectation that there is an interaction between the students' pattern and level of proficiency on one hand, and the choice of examination subjects on the other hand. Therefore, the third model has a multidimensional representation of proficiency where the choice-variables are explicitly modeled.

Model 1. Model 1 is the unidimensional version of the generalized partial credit model (Muraki, 1992). The probability that the grade X_{ni} is in category j ($j = 0, \dots, m$) is given by

$$p\,(X_{ni} = j \mid d_{ni} = 1; \theta_n) = \frac{\exp\left(j\alpha_i\theta_n - \sum_{h=1}^{j}\beta_{ih}\right)}{1 + \sum_{h=1}^{m}\exp\left(h\alpha_i\theta_h - \sum_{k=1}^{h}\beta_{ik}\right)}, \tag{3}$$

where θ_n is the unidimensional proficiency parameter that represents the overall proficiency of student n. So it is assumed here that all examination grades relate to one unidimensional proficiency parameter θ_n. The parameters $\beta_{ij}(j = 1, \dots, m)$ are the locations on the latent scale where the probabilities of scoring in category $j - 1$ and j are equal. These parameters model the difficulty of examination subject i ($\beta_{i0} = 0$ to identify the model). Parameter α_i defines the extent to which the response is related to the proficiency θ_n.

The parameters of the model can be estimated using marginal maximum likelihood (MML; see Bock & Aitkin, 1981). In MML; the model is enhanced with the assumption that the proficiency parameters are drawn from one or more normal distributions (the latter is known as multiple-group IRT; see Bock & Zimowski, 1997). In the example presented below, it cannot be assumed a priori that the average level of proficiency is independent of the chosen examination package. Therefore, it will be assumed that students choosing the same examination package (that is, students with the same pattern on the choice variables $d_{n1}, \ldots, d_{ni}, \ldots, d_{nK}$) are drawn from a normal distribution with a mean μ_p (where p is the index of the package) and a variance σ^2.

In MML, a likelihood function is maximized where the students' proficiency parameters are integrated out of the likelihood. The marginal log-likelihood for Model 1 is given by

$$L_1 = \sum_p \sum_{n\,|\,p} \log \int \prod_i p(x_{ni}\,|\,d_{ni};\theta) g(\theta; \mu_p, \sigma^2)\, d\theta, \tag{4}$$

where x_{ni} is the observed grade, $p(x_{ni} \mid d_{ni};\, \theta)$ is equal to equation (3) evaluated at x_{ni} if $d_{ni} = 1$, and $p(x_{ni} \mid d_{ni};\, \theta) = 1$ if $d_{ni} = 0$. Further, $g(\theta;\, \mu_p,\, \sigma^2)$ is the normal density with parameters μ_p and σ^2. The model can be identified by choosing $\mu_1 = 0$ and $\sigma^2 = 1$.

The estimates can be computed using the software packages MULTILOG (Thissen, Chen, & Bock, 2002) or PARSCALE (Muraki & Bock, 2002). These packages compute concurrent MML estimates of all the free structural parameters in the model (the α- and β-parameters and the means μ_p for $p > 1$). This is the approach that is also pursued in the present article.

After the parameters of the examinations are estimated by MML, the missing examination scores can be estimated by their posterior expectations, that is, by

$$E\left(X_{ni} \mid \mathbf{x}_n\right) = \sum_{j=1}^{m} j \int p(X_{ni} = j \mid d_{ni} = 1; \theta) p(\theta \mid \mathbf{x}_n)\, d\theta, \tag{5}$$

where $p(\theta \mid \mathbf{x}_n)$ is the distribution of θ given the observations $\mathbf{x}_n$, and $p(X_{ni} = j \mid d_{ni} = 1;\, \theta)$ is defined by equation (3). These expected scores are then used in equation (2).

Model 2. In the previous model; it was assumed that the grade of student n depended on a unidimensional proficiency parameter θ_n. However,there may be more than one proficiency factor underlying the grades. For instance, there might be a specific proficiency factor for the science subjects and another one for the language subjects. If Q proficiency dimensions are needed to model the grades, the proficiency can be represented by a vector of proficiency parameters $\boldsymbol{\theta}_n = (\theta_{n1}, \ldots, \theta_{nq}, \ldots, \theta_{nQ})$. The probability of a grade in category j is now given by

$$p\left(X_{ni} = j \mid d_{ni} = 1; \boldsymbol{\theta}_n\right) = \frac{\exp\left(j\left(\sum_{q=1}^{Q} \alpha_{iq}\theta_{nq}\right) - \sum_{h=1}^{j} \beta_{ih}\right)}{1 + \sum_{h=1}^{m} \exp\left(h\left(\sum_{q=1}^{Q} \alpha_{iq}\theta_{nq}\right) - \sum_{k=1}^{h} \beta_{ik}\right)}. \tag{6}$$

In addition, it will be assumed that the proficiency parameters $\boldsymbol{\theta}_n$ of groups of students taking a specific package of examination subjects have a Q-variate normal distribution with a mean $\boldsymbol{\mu}_p$ and a covariance matrix $\boldsymbol{\Sigma}$. So it is assumed that the mean depends on the examination package and that the covariance matrix of the proficiency parameters is common for all students. Takane and de Leeuw (1987) show that the model is equivalent to a full-information factor analysis model. Therefore, the parameters $\alpha_{i1}, \ldots, \alpha_{iQ}$, which are usually referred to as discrimination parameters, but may also be called factor loadings, and the proficiency parameters $\theta_{n1}, \ldots, \theta_{nq}, \ldots, \theta_{nQ}$ can be viewed as factor scores. Note that the factor loadings are specific for an examination subject and that they model the relation between the probability of obtaining a grade and the level on the Q proficiency dimensions. A high positive value of α_{iq} means that the q-th dimension is important for the subject, and a value close to zero means that the dimension does not play an important role.

The model is identified by setting $\boldsymbol{\mu}_1 = 0$ and setting the diagonal of $\boldsymbol{\Sigma}$ equal to one. For a discussion of these and alternative identification restrictions refer to Béguin and Glas (2001). The marginal log-likelihood of the model becomes

$$L_2 = \sum_{p} \sum_{n \mid p} \log \int .. \int \prod_{i} p(x_{ni} \mid d_{ni}; \boldsymbol{\theta}) g(\boldsymbol{\theta}; \boldsymbol{\mu}_p, \Sigma) \, d\boldsymbol{\theta}, \tag{7}$$

where x_{ni} is the observed grade, $p(x_{ni} \mid d_{ni}; \boldsymbol{\theta})$ is equal to equation (6) evaluated at x_{ni} if $d_{ni} = 1$, and $p(x_{ni} \mid d_{ni}; \boldsymbol{\theta}) = 1$ if $d_{ni} = 0$, and $g(\boldsymbol{\theta}; \boldsymbol{\mu}_p, \boldsymbol{\Sigma})$ is the Q-variate normal density. The parameters of the multidimensional model can be estimated using MML (Bock, Gibbons, & Muraki, 1988) and the computer packages TESTFACT (Wood et al., 2002), ConQuest (Wu, Adams, & Wilson, 1997) or Mplus (Muthén & Muthén, 2003) can be used to compute the estimates.

Using the MML estimates of the parameters of the examinations, the missing examination scores can be estimated by their posterior expectations analogously to (5), but the expectations are now with respect to a Q-variate posterior distribution $p(\boldsymbol{\theta} \mid \boldsymbol{x}_n)$, so

$$E\,(X_{ni} \mid \mathbf{x}_n) = \sum_{j=1}^{m} j \int .. \int p(X_{ni} = j \mid d_{ni} = 1; \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \mathbf{x}_n) \, d\boldsymbol{\theta}. \tag{8}$$

Model 3. In Model 2 there is no interaction between the choice of an examination subject and the proficiency parameters. That is, it is assumed that the process causing the missing data does not need to be considered in the estimation process. Rubin (1976) identified two conditions under which the missing data process can be ignored. A missing data mechanism is ignorable if the missing values are missing at random (MAR) and if the parameters of the distribution of the observed data (say λ) and the distribution of the missing data (say φ) are distinct. MAR holds if the probability of the missing data pattern $p(d \mid x_{mis}, x_{obs}, \varphi)$ does not depend on missing data, that is, if $p(d \mid x_{mis}, x_{obs}, \varphi) = p(d \mid x_{obs}, \varphi)$. Distinctness entails that there are no functional dependencies between φ and λ, or that φ and λ have independent priors. If ignorability does not hold, the inferences made using an IRT model ignoring the missing data process can be severely biased (Bradlow & Thomas, 1998; Holman & Glas, 2005).

A general method to deal with nonignorable missing data proposed by Heckman (1979) is the introduction of a selection model. Several authors have applied this approach in the framework of IRT (Holman & Glas, 2005; Moustaki & Knott, 2000; Moustaki & O'Muircheartaigh, 2000) and have shown that selection bias can be removed when the distribution of d_{ni} is modeled concurrently with the observed data using an IRT model. Their approach is adapted to the present problem as follows. As in Model 2, the scores on the examination subjects are modeled by equation (6). Further, it is assumed that there exists a latent variable θ_{Q+1} that governs the choice of the examination subjects, that is, the realizations of the choice variable defined by equation (1). If the students' proficiency level is highly correlated with the choice of examination subjects, then θ_{Q+1} will be highly correlated with $\theta_1, \ldots,$ θ_Q also. The dependence between the latent variables is modeled by assuming that $\theta_1, \ldots, \theta_{Q+1}$ have a multivariate normal distribution, again with a specific mean for every group of students and a common covariance matrix. The marginal likelihood of the model is

$$L_3 = \sum_p \sum_{n \mid p} \log \int .. \int \prod_i \left[p(x_{ni} \mid d_{ni}; \boldsymbol{\theta}^{(Q)}) p(d_{ni}; \theta_{Q+1}) \right] g(\boldsymbol{\theta}; \boldsymbol{\mu}_p, \Sigma) \, d\boldsymbol{\theta}, \quad (9)$$

where $\boldsymbol{\theta}^{(Q)} = (\theta_1, \ldots, \theta_Q)$, and $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_{Q+1})$.

The correlation between θ_{Q+1} and the proficiency dimensions $\theta_1, \ldots, \theta_Q$ describe the extent to which the choice of an examination subject depends on the proficiency level. The magnitude of the correlations between $\theta_1, \ldots, \theta_Q$, and θ_{Q+1} gives an indication of the extent to which ignorability is violated. If these correlations are close to zero, the choice behavior is not related to proficiency, and the missing data are ignorable. If, on the other hand, these correlations are substantial the choice variable is highly related to the proficiency for the students. Holman and Glas (2005) show that the bias in the parameter estimates is positively related to the correlation between the latent proficiency and the parameters of the IRT model for the missing data indicator d_{ni} and that this bias vanishes when the observations and the realizations of the missing data indicators are concurrently modeled by the multidimensional IRT model described here.

The final consideration is about the model for the choice variables d_{ni} . Since the students can only choose a limited number of subjects, it is reasonable to assume that the probability of choosing a subject as a function of the proficiency dimension θ_{Q+1} is single peaked: Students will probably choose subjects within a certain region of the proficiency dimension θ_{Q+1} and avoid subjects that are too difficult or too easy. An IRT choice model that may reflect this is given by

$$p\,(d_{ni} = 1) = \pi_{i1}(\theta_{(Q+1)n}) - \pi_{i2}(\theta_{(Q+1)\,n}), \quad (10)$$

where

$$\pi_{ij}(\theta_{(Q+1)\,n}) = \frac{\exp\left(\theta_{(Q+1)\,n} - \gamma_{ij}\right)}{1 + \exp\left(\theta_{(Q+1)\,n} - \gamma_{ij}\right)}, \tag{11}$$

and $\gamma_{i1} < \gamma_{i2}$ to guarantee that $\Pr(d_{ni} = 1)$ is positive. The model given by equation (10) is related to a special case of the graded response model by Samejima (1969, 1973). The graded response model pertains to a polytomously scored response variable, for instance, a response variable y_{ni} that assumes the values 0, 1 or 2. The response probabilities are given by

$$\Pr(y_{ni} = 0) = 1 - \pi_{i1}(\theta)$$

$$\Pr(y_{ni} = 1) = \pi_{i1}(\theta) - \pi_{i2}(\theta)$$

$$\Pr(y_{ni} = 2) = \pi_{i2}(\theta),$$

with $\pi_{ij}(\theta)$ as defined by equation (11). So model equation (11) can be derived from the graded response model by noting that the responses $y_{ni} = 0$ and $y_{ni} = 2$ are extreme cases and collapsed to $d_{ni} = 0$.

The model considered here is closely related to models by Verhelst and Verstralen (1993) and Andrich and Luo (1993; also see Andrich, 1997). Also, these two models have a single-peaked response probability, but the functional form of the probability is chosen differently. The model by Verhelst and Verstralen is a collapsed version of the partial credit model; the model by Andrich and Luo has a hyperbolic cosine function probability function. The motivation for the present model is its simple functional form.

Estimation procedures for a model that is a mixture of the logistic IRT model defined by equation (6), the collapsed graded response model defined by equation (10), and a $(Q + 1)$-variate normal model for the proficiency parameters are not readily available. The MML procedure used to calculate the estimates reported below is outlined in Glas and Korobko (2007). Estimation of the missing examination scores is analogous to their estimation in Model 1 and Model 2, except that the expectations are now with respect to a $Q + 1$-variate posterior distribution $p(\boldsymbol{\theta} \mid \mathbf{x}_n)$.

Model Fit

Likelihood ratio testing is the standard methodology for model comparison and this methodology will also be applied below. However, these tests are rather global and give no information with respect to the fit of specific examination subjects. In principle, IRT models can be evaluated by Pearson-type statistics, that is, statistics based on the difference between observations and their expectations under the null-model. Such statistics are available for unidimensional models for dichotomous observations (Glas & Suarez-Falcon, 2003; Orlando & Thissen, 2000), for unidimensional models for polytomous observations (Glas, 1998, 1999), and for multidimensional models for such observations (te Marvelde, Glas, Van Landeghem,

& Van Damme, 2006). In the present article, a comparable fit statistic will be presented that is targeted at the special application considered here.

Most item fit statistics are based on splitting up the sample of respondents into subgroups with different proficiency distributions and evaluating whether the item response frequencies in these subgroups differ from their expected values. Orlando and Thissen (2000) point out that the splitting criteria should be directly observable (for instance, number-correct scores) rather than estimated (for instance, estimated proficiency parameters). Following this suggestion, we split up the sample of students using a splitter examination labeled s. Two subgroups are formed, one subgroup of students that did choose subject s (so $d_{ns} = 1$) and one subgroup of students that did not choose subject s (so $d_{ns} = 0$). The test is based on the assumption that this criterion splits the sample up in two subgroups with different proficiency distributions. We compute the average grade on subject i of students in both subgroups as

$$S_{i0} = \left[\sum_n (1 - d_{ns})\, d_{ni} x_{ni}\right] \Big/ \left[\sum_n (1 - d_{ns})\, d_{ni} m\right]$$

and

$$S_{i1} = \left[\sum_n d_{ns}\, d_{ni} x_{ni}\right] \Big/ \left[\sum_n d_{ns}\, d_{ni} m\right],$$

where m is the maximum grade on examination i. These average grades can be compared with their expected values given by

$$E_{i0} = \left[\sum_n (1 - d_{ns})\, d_{ni} E\left(X_{ni} \mid \mathbf{x}_n\right)\right] \Big/ \left[\sum_n (1 - d_{ns})\, d_{ni} m\right]$$

and

$$E_{i1} = \left[\sum_n d_{ns}\, d_{ni} E\left(X_{ni} \mid \mathbf{x}_n\right)\right] \Big/ \left[\sum_n d_{ns}\, d_{ni} m\right],$$

where $E\left(X_{ni} \mid \mathbf{x}_n\right)$ is given by equation (8).

Glas and Korobko (2007) show that a Pearson-type fit statistic based on the squared differences between observed and expected values can be used to evaluate whether the observed and expected response frequencies are acceptably close given the observed value on the choice variable of the splitter examination. They also outline that the statistic has an asymptotic χ^2 distribution with one degree of freedom. However, the application presented below has a very large sample size and the power of the test becomes very large. Therefore, the test will be used to compare the relative model fit of nested models. More specifically, the test will be used to evaluate whether Model 3 fits the data systematically better than Model 2.

TABLE 1
Distribution of Students Over Examination Subjects in Original Data Set (N = 16,118) and Analyses Data Set (N = 6,142)

Subjects Selected	Percentage Original Data	Percentage Selected Data	Subjects Not Selected	Percentage Original Data
Dutch Language	99.9	100.0	Frisian Language	.0
Latin	14.6	10.3	Russian	.0
Classical Greek	6.2	4.1	Spanish	.2
French	37.6	36.6	Handicrafts	1.9
German	45.4	44.5	Music	1.6
English	99.1	100.0	Philosophy	.7
History	49.5	48.8	Social studies	2.3
Geography	33.9	31.3		
Applied Math	63.0	65.1		
Advanced Math	44.7	40.2		
Physics	46.7	42.5		
Chemistry	38.2	39.9		
Biology	37.0	33.3		
General Economy	58.7	66.6		
Business Economy	36.0	37.9		
Art	7.8	5.5		

An Example

The Data

The data used to illustrate the methods are from approximately 18-year-old students in the preuniversity track in the Netherlands. This track gives direct entry into a university. The examinations are standardized nationwide achievement tests. The students take examinations in 7 or 8 subjects chosen from the list of subjects displayed in Table 1. The data used in this study were collected by the Dutch Inspection of Education. For this study, only the results from the first session of the examinations were used (unsatisfactory marks might be "repaired" in a re-session). The data are a subset of the data of preuniversity students that took their final examination in the school year 1994/1995. The original data set comprised 16,118 students. To keep the presentation of the results relatively simple, the analysis was restricted to 60 fairly common combinations of examination subjects. The resulting data set consisted of the examination results of 6,142 students. The distribution of the students over examination subjects in the original data and the selected data are shown in Table 1.

Below, the appropriateness of the methods for computing adjusted GPAs are assessed by evaluating the consequences of the method in subgroups. From the combinations of different subjects chosen by the students, we distinguish three main groups:

1. The language-oriented students (20%). These students take examinations in French and German languages, and examinations in not more than one of the subjects Applied Mathematics, Advanced Mathematics, Physics, and Chemistry.

TABLE 2
Observed Examination Scores Per Subject and Per Package

Subjects	Overall	Science	Language	Mixed
Dutch Language	1.38	1.29	1.53	1.37
Latin	2.47	2.36	2.44	2.70
Classical Greek	2.18	2.07	2.18	2.32
French	1.63	—	1.68	1.57
German	1.50	—	1.51	1.50
English	1.50	1.38	1.64	1.51
History	1.58	1.87	1.52	1.55
Geography	1.31	1.88	1.11	1.34
Applied Math	1.15	2.28	.71	.91
Advanced Math	1.37	1.37	—	1.36
Physics	1.50	1.47	—	1.59
Chemistry	1.76	1.76	—	1.75
Biology	1.75	1.71	—	1.91
General Economy	1.27	1.73	.92	1.25
Business Economy	1.41	1.84	1.39	1.30
Art	1.60	1.75	1.56	1.57

2. The science-oriented students (33%). These students take examinations in at least three of the subjects Applied Mathematics, Advanced Mathematics, Physics, and Chemistry, and no examinations in French or German languages.
3. All other students (47%).

The original grades ranged from 1 (“poor”) to 10 (“excellent”), but for the purpose of our study these were rescaled to a four point scale, where the points are 0 (original grade 0 to 5.4, which is unsatisfactory), 1 (original grade 5.5 to 6.4, which is just satisfactory), 2 (original grade 6.5 to 7.4, which is good), and 3 (original grade 7.5 to 10, which is very good). The overall observed mean examination scores and the mean examination scores observed in the three subgroups are displayed in Table 2.

The following observations are of interest. Note that students with a science-oriented package score lower on Dutch and English language than the students with a language-oriented package. On the other hand, students with a science-oriented package score substantially higher on Applied Mathematics than students with a language-oriented package. This is a first indication that the proficiency dimension might not be unidimensional. Further, the overall mean scores on French and German languages may be boosted relative to the overall mean scores on Dutch and English languages by the absence of the students with a science-oriented package, who seem to have a lower language proficiency than the other students. The IRT analyses presented below will clarify these observations.

Results

Model 1. Model 1 was estimated by MML, that is, by maximizing equation (4). The estimates of the parameters α_i and $\beta_{ij}(j = 1, \ldots, m)$ are given in Table 3. The last

TABLE 3
Parameter Estimates for Model 1

Subjects	N	α	β_1	β_2	β_3	$\bar{\beta}$
Dutch Language	6142	.38	−.97	.10	1.38	.17
Latin	637	.65	−1.95	−1.18	−.17	−1.10
Classical Greek	256	.68	−1.36	−.65	.57	−.48
French	2250	.91	−1.24	−.28	.71	−.27
German	2739	.99	−1.22	−.05	.86	−.14
English	6142	.63	−.79	−.03	.79	−.01
History	2997	.83	−1.31	−.24	.92	−.21
Geography	1928	.71	−1.22	.14	1.41	.11
Applied Math	4002	.59	−.42	.33	1.01	.31
Advanced Math	2471	.91	−.51	.44	1.23	.39
Physics	2614	1.38	−.90	.29	1.51	.30
Chemistry	2452	1.37	−1.14	−.15	.97	−.11
Biology	2048	1.03	−1.77	−.25	1.35	−.22
General Economy	4092	.87	−.82	.23	1.30	.24
Business Economy	2330	.90	−1.10	−.01	1.21	.03
Art	338	.37	−1.60	−.22	1.06	−.26

column of the table gives the average of the estimates of the parameters $\beta_{ij}(j = 1, \ldots, m)$, denoted by$\bar{\beta}$. This average is an estimate of the global position of subject j on the latent scale, and serves as an indication of the average difficulty of the subject.

Examination of the α estimates shows that Dutch language and Art are the least discriminating with respect to the overall proficiency and Physics and Chemistry have the highest discrimination. Inspection of the values of $\bar{\beta}$ in the last column shows that Advanced Mathematics is now slightly more difficult than Applied Mathematics. This result is contrary to the result in Table 2, where the overall average of Advanced Mathematics is higher than the overall average of Applied Mathematics (1.37 versus 1.15). This phenomenon is of course explained by the fact that the students with a language-oriented package do not take Advanced Mathematics.

Using the MML estimates, posterior expectations as defined by Equation (5) were imputed for the missing examination scores. The results are given in Table 4. The average scores in the table can be interpreted as the average scores obtained if all students endorsed all subjects. The most dramatic effect is the decrease of the average scores for the classical languages Latin and Greek. The explanation may be that the small percentage of the students that actually choose these subjects (10.3% and 4.1%) are highly proficient. Adding imputed values for the other students (of lower proficiency) can only lower this average. This explanation is in line with the experience in Dutch education.

Unexpected results in Table 4 were the imputed means for French and German languages for the students with a science-oriented package. In Table 2, it can be verified that these students did not choose these two languages in their examination package. In Table 2, it can be seen that these students scored relatively low on Dutch and

TABLE 4
Examination Scores Per Subject and Per Package Estimated Under Model 1

Subjects	Science	Language	Mixed
Dutch Language	1.29	1.53	1.37
Latin	1.90	1.86	1.83
Classical Greek	1.61	1.58	1.57
French	1.58	1.68	1.56
German	1.55	1.51	1.51
English	1.38	1.64	1.51
History	1.63	1.53	1.55
Geography	1.53	1.31	1.42
Applied Math	1.76	.87	1.01
Advanced Math	1.37	1.37	1.38
Physics	1.47	1.41	1.46
Chemistry	1.74	1.46	1.54
Biology	1.67	1.48	1.56
General Economy	1.57	1.04	1.28
Business Economy	1.58	1.44	1.38
Art	1.69	1.61	1.63

English languages (1.29 and 1.38, respectively), yet the imputed means on French and German languages are quite close to the mean scores for the students with a language-oriented package. The opposite phenomenon occurred with the imputed values for Advanced Mathematics, Physics, Chemistry, and Biology for the students with a language-oriented package. The imputed means were all close to the means for the other students, yet in Table 4 it can be seen that their (generally observed) score on Applied Mathematics was as low as .87. This is highly unexpected. In the sequel, it will become clear that this phenomenon is attributable to the multidimensionality of the proficiency variables.

Model 2. A three-dimensional version of Model 2 was fitted with a method by Béguin and Glas (2001). The method identifies the dimensions by fitting unidimensional IRT models. Items, or, in the present case, examination subjects, are discarded from a scale until a unidimensional IRT model that fits the data is found. The process is then repeated using the discarded items. The unidimensional scales identified in this manner are entered as unique factors in a multidimensional model. So most of the examination subjects load on one dimension only. Examination subjects that do not fit any dimension are allowed to load on all dimensions. In the present application, the unidimensional subscales were searched for with the program OPLM (Verhelst, Glas & Verstralen, 1995). The R_{1c} statistic (Glas, 1988) was used as a criterion for model fit. Finally, given this structure, an MML estimate was made of the subject parameters and the correlation matrix.

The results are shown in Table 5 under the heading “Factor Solution Model 2.” The factor loadings fixed to zero are marked by an asterisk. The three dimensions that appeared can be interpreted as “Language,” “Science,” and “Economy.” Examination

TABLE 5
Factor Loading Per Subject for the Three- and Four-Factor Solution and Correlation Matrices

	Factor Solution Model 2			Factor Solution Model 3			
Subjects	Language	Science	Economy	Language	Science	Economy	Choice
Dutch Language	2.22	−.05	.45	1.91	−.09	.44	–
English	6.97	.00*	.00*	5.46	.00*	.00*	–
Latin	3.14	−.22	1.89	2.88	−.32	1.67	−.76
Classical Greek	2.31	.03	2.75	2.32	−.20	1.46	−1.12
French	7.26	.00*	.00*	6.11	.00*	.00*	−.89
German	9.27	.00*	.00*	7.79	.00*	.00*	−.62
History	2.04	1.31	2.06	1.86	−.23	2.18	−.19
Geography	.00*	6.12	.00*	.00*	3.23	.00*	.26
Applied Math	.00*	3.63	.00*	.00*	4.69	.00*	.01
Advanced Math	−.76	5.84	.09	−.64	4.25	.12	.43
Physics	.00*	9.03	.00*	.00*	6.01	.00*	.76
Chemistry	.00*	8.86	.00*	.00*	6.57	.00*	.89
Biology	.00*	6.62	.00*	.00*	5.09	.00*	1.24
General Economy	.00*	.00*	7.78	.00*	.00*	3.42	−.31
Business Economy	.00*	.00*	6.99	.00*	.00*	4.26	−.13
Art	1.23	.03	1.06	1.15	.08	.49	.56
Correlation Matrix							
Language	1.00			1.00			
Science	.51	1.00		.43	1.00		
Economy	.45	.81	1.00	.48	.84	1.00	
Choice dimension				.12	.74	.56	1.00

*Fixed factor loadings.

subjects that do not completely load according to common expectations are Latin and Classical Greek (both load high on the Language and Economy dimension). Note that History loads on all three dimensions and Geography has a high loading on the Science dimension. The correlation matrix of the three latent dimensions is shown at the bottom of the table. Note that the correlation between the Science and Economy dimensions is substantially higher than the other two correlations.

Next, using the MML estimates of the parameters of Model 2, the missing scores were estimated by their posterior expected values. The results are given in Table 6 under the heading Model 2. Again, the average scores in the table can be interpreted as the average scores obtained if all students endorsed all subjects. An important implausible finding using Model 1 was that the expected grades for the language-oriented group on Advanced Mathematics, Physics, Chemistry, and Biology were almost as high as the grades of the science-oriented group. Inspection of the analogous estimated averages computed using Model 2 displayed in Table 6 shows that these estimates do not suffer from this phenomenon. Also, the estimates for French and German languages for the science-oriented group are now lower than the analogous estimates in Table 4.

TABLE 6
Examination Scores Per Subject and Per Package Estimated Under Model 2 and Model 3

	Model 2			Model 3		
Subjects	Science	Language	Mixed	Science	Language	Mixed
Dutch Language	1.29	1.53	1.37	1.29	1.53	1.37
Latin	1.80	1.78	1.78	1.85	1.79	1.83
Classical Greek	1.54	1.48	1.56	1.49	1.62	1.64
French	1.45	1.68	1.54	1.44	1.69	1.58
German	1.41	1.51	1.49	1.35	1.50	1.51
English	1.38	1.64	1.51	1.38	1.64	1.51
History	1.66	1.52	1.59	1.72	1.62	1.69
Geography	1.88	1.26	1.50	1.85	1.11	1.41
Applied Math	1.97	.88	1.09	1.86	.82	1.03
Advanced Math	1.36	.81	1.03	1.36	.81	1.17
Physics	1.47	.92	1.15	1.46	.75	1.04
Chemistry	1.75	1.00	1.25	1.75	.85	1.17
Biology	1.69	1.06	1.31	1.72	.97	1.28
General Economy	1.50	1.03	1.31	1.45	1.02	1.24
Business Economy	1.49	1.29	1.37	1.43	1.22	1.30
Art	1.64	1.59	1.65	1.57	1.87	1.85

Model 3. Above, it was argued that the free choice of examination subjects leads to a stochastic design that might violate the assumption of ignorability. Therefore, Model 3 was derived from Model 2 by adding a special dimension to model the choice variables d_{ni} defined in equation (1). The MML estimates of the parameters of Model 3 were obtained by maximization of equation (9); the results are shown in Table 5 under the heading “Factor Solution Model 3.” Note that the patterns of the factor loadings and the correlation matrices for the first three dimensions for Model 2 and Model 3 are similar. Interestingly, the choice dimension correlates highest with the Science dimension. So the choice students make is mostly related to their Science ability.

Displaying the factor loadings of the subjects on the choice dimension is not very informative since they are all equal to one. Therefore, the average of the two subject parameters, that is, $\bar{\gamma}_i = (\gamma_{i1} + \gamma_{i2})/2$ are displayed for all subjects in the last column labeled “Choice.” The parameters $\bar{\gamma}_i$ can be seen as an estimate of the location of the subject on this fourth proficiency dimension. Note that the parameters for Dutch and English cannot be estimated, because these two examination subjects are obligatory and so all the choice variables d_{ni} for these examination subjects are structurally equal to one.

The interpretation of the mean parameters $\bar{\gamma}_i$ is as follows. The fourth dimension correlates positively with the three proficiency dimensions, and highest with the Science dimension. This dimension can be viewed as an overall proficiency dimension, and the choice of subjects is assumed governed by proficiency. Since the difficulty parameters $\bar{\gamma}_i$ are an estimate of the location of the subjects on the fourth proficiency dimension, they represent the ordering of the examination subjects on this dimension. That is, such difficult subjects as Biology ($\bar{\gamma}_i = 1.24$), Chemistry ($\bar{\gamma}_i = 0.89$) and Advanced Mathematics ($\bar{\gamma}_i = 0.43$) are endorsed by the more proficient students. Note also that Art ($\bar{\gamma}_i = 0.56$) scores high on this dimension.

As for the other two models, for Model 3 the missing scores were estimated by their posterior expectations. The averages computed assuming all students endorsed all subjects are given in Table 6 under the heading Model 3. Also the estimates under Model 3 do not show the implausible results obtained under Model 1. In Table 6, it can also be seen that the estimates for Model 2 and Model 3 did not substantially differ.

Model Fit

First, the fit of the three models was compared using likelihood ratio tests. A test of Model 1 against Model 2 yielded a chi-square value of 2070.1 with 135 degrees of freedom. So Model 1 was rejected. To facilitate the test of Model 3 against Model 2, both models have to refer to the same data. For Model 3, these data comprise subject scores and choice variables. Therefore, Model 2 was enhanced with the choice model defined in equation (11). Then the likelihood was computed as the product of the likelihood under Model 2 multiplied by the likelihood of the choice model for the variables d_{ni}. The test of this enhanced model against Model 3 is equivalent to testing whether the covariances between the latent variables associated with the observations x_{ni} and the latent variables associated with d_{ni} are zero. The test statistic has a value of 312.2, with 3 degrees of freedom. The conclusion is that Model 3 fitted significantly better than Model 2. However, as noted above, the impact of this better model fit was quite small.

Likelihood ratio tests are global tests that give an impression of overall model fit. They do not provide information on the fit of the individual examination subjects. Therefore, the fit statistic as defined above was computed for all examinations $i = 1, \ldots, 16$, both under Model 2 and Model 3. Three splitter examinations were used: Advanced Mathematics, History, and Business Economy. The results are displayed in Table 7. Above it was argued that the absolute values of the test statistics were less interesting due to the large sample sizes. The statistics have an asymptotic χ^2 distribution with one degree of freedom, and in Table 7 it can be seen that most are significant (the 5% critical value is 3.84). More informative for the comparison of the two models is their difference in model fit. In Table 7, all instances where Model 3 fitted better than Model 2 are marked with an asterisk. In 36 of the 45 cases, Model 3 fitted better than Model 2. So also here the overall conclusion is that Model 3 showed the best fit.

Discussion and Conclusion

The problem addressed concerns the comparison of examination grades in cases where students have chosen different subjects. The complicating factor is that students only sit examinations in subjects they have chosen themselves. As a consequence, more proficient students may choose examinations in subjects that are more difficult and the less bright students may choose less difficult subjects. However, it is not a priori plausible that the proficiency structure assessed is unidimensional. This

TABLE 7
Model Fit Evaluated UsingT_i-Statistic

Splitter:	Advanced Math		History		Business Economy	
Subjects	Model 2	Model 3	Model 2	Model 3	Model 2	Model 3
Dutch Language	84.0	31.0*	58.9	9.3*	106.8	38.3*
Latin	9.0	1.3*	12.0	6.4*	13.3	4.4*
Classical Greek	.7	1.2	1.0	.2*	1.0	1.1
French	2.1	0.0	25.2	15.4*	41.3	34.2*
German	.1	1.5	52.9	47.7*	63.0	50.2*
English	48.9	34.9*	69.4	48.8*	125.1	98.8*
History	1.5	1.4*			79.8	56.4*
Geography	4.4	2.5*	176.5	113.0*	89.4	33.0*
Applied Math	25.7	70.6	17.8	20.2	90.0	35.2*
Advanced Math			22.7	6.9*	17.4	21.1
Physics	19.7	4.2*	10.5	2.4*	23.7	12.2*
Chemistry	67.3	22.8*	22.6	6.1*	.9	.4*
Biology	7.4	5.3*	125.3	117.8*	7.3	4.9*
General Economy	9.4	4.5*A	91.4	49.0*		
Business Economy	1.8	1.2*	4.0	2.6	2.6	5.5
Art	11.5	9.4*	13.4	13.6	15.0	9.0*

*Indicates better fit for Model 3.

was corroborated by the implausible result that the language-oriented students had almost as high expected grades in Advanced Mathematics, Physics, and Chemistry as the science-oriented students when a unidimensional model was used to compute these expectations. Therefore, a multidimensional IRT model for polytomous items, the generalized partial credit model, was fitted to the data. The three-dimensional model had a substantially better fit than the unidimensional model. Furthermore, the implausible result of the high expected grades in Mathematics and Physics for the language-oriented students under a unidimensional model now vanished.

Another problem addressed related to the fact that it is not a priori plausible that the missing data (the grades for the examination subjects that were not chosen) are missing at random. In other words, it was expected that the missingness indicators correlated with the proficiency level in such a way that this might bias the estimates of the difficulty of the examination subjects. It was attempted to remove this bias by using a four-dimensional IRT model, where the first three dimensions are related to the observed grades, while the fourth dimension is related to the choice variables. Though this model fitted the data significantly better than the three-dimensional model, the expected grades computed using the two models were very similar.

The models presented here were exemplified using examination data. However, they are also relevant in other situations where grade point averages play a role. The most sophisticated method to communicate the proficiency level of the students would of course be to report their multidimensional IRT proficiency parameters.

Such a report would, however, be unintelligible for most audiences. In the present article, we show that adjusted GPAs can be computed that take the difficulty of the subjects, multidimensionality, and choice behavior into account.

References

Andrich, D. (1997). A hyperbolic cosine IRT model. In W. J. van der Linden and R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 399–414). New York: Springer.

Andrich, D., & Luo, G. Z. (1993). A hyperbolic cosine latent trait model for unfolding dichotomous single-stimulus responses. *Applied Psychological Measurement, 17,* 253–276.

Béguin, A. A., & Glas, C. A. W. (2001). MCMC estimation and some fit analysis of multidimensional IRT models. *Psychometrika, 66,* 541–562.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: An application of an EM-algorithm. *Psychometrika, 46,* 443–459.

Bock, R. D., Gibbons, R. D., & Muraki, E. (1988). Full-information factor analysis. *Applied Psychological Measurement, 12,* 261–280.

Bock, R. D., & Zimowski, M. F. (1997). Multiple group IRT. In W. J. van der Linden and R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 433–448). New York: Springer.

Bradlow, E. T., & Thomas, N. (1998). Item response theory models applied to data allowing examinee choice. *Journal of Educational and Behavioral Statistics, 23,* 236–243.

Caulkins, J. P., Larkey, P. D., & Wei, J. (1996). *Adjusting GPA to reflect course difficulty.* Pittsburgh, PA: Carnegie Mellon University, The Heinz School of Public Policy and Management.

Fitz-Gibbon, C. T. (1994). *Monitoring education: Indicators, quality and effectiveness.* London: Cassell.

Glas, C. A. W. (1988). The derivation of some tests for the Rasch model from the multinomial distribution. *Psychometrika, 53,* 525–546.

Glas, C. A. W. (1998). Detection of differential item functioning using Lagrange multiplier tests. *Statistica Sinica, 8,* 647–667.

Glas, C. A. W. (1999). Modification indices for the 2-pl and the nominal response model. *Psychometrika, 64,* 273–294.

Glas, C. A. W., & Korobko, O. B. (2007). *A marginal maximum likelihood procedure for an IRT model with single-peaked response functions.* OMD Progress Report, 07–01. University of Twente, Univerity the Netherlands. Available on www.utwente.nl.

Glas, C. A. W., & Suarez-Falcon, J. C. (2003). A comparison of item-fit statistics for the three-parameter logistic model. *Applied Psychological Measurement, 27,* 87–106.

Heckman, J. (1979). Sample selection bias as a specification error. *Econometrika, 46,* 153–161.

Holman, R., & Glas, C. A. W. (2005). Modelling non-ignorable missing data mechanisms with item response theory models. *British Journal of Mathematical and Statistical Psychology, 58,* 1–17.

Johnson, V. E. (1997). An alternative to traditional GPA for evaluating student performance. *Statistical Science, 12,* 251–278.

Johnson, V. E. (2003). *Grade inflation: A crisis in college education*. New York: Springer.

Kelly, A. (1976). A study on the comparability of external examinations in different subjects. *Research in Education, 1,* 37–63.

Linn, R. L. (1966). Grade adjustments for prediction of academic performance: A review. *Journal of Educational Measurement, 3,* 313–329.

Moustaki, I., & Knott, M. (2000). Weighting for item non-response in attitude scales by using latent variable models with covariates. *Journal of the Royal Statistical Society, Series A, 163,* 445–459.

Moustaki, I., & O'Muircheartaigh, C. (2000). A one dimensional latent trait model to infer attitude from nonresponse for nominal data. *Statistica, 2,* 259–276.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16,* 159–176.

Muraki, E., & Bock, R. D. (2002). *PARSCALE: Parameter scaling of rating data* [Computer Program and Manual]. Chicago: Scientific Software.

Muthén, L. K., & Muthén, B. O. (2003). *Mplus* [Computer Program and Manual]. Los Angeles: Muthén & Muthén.

Newton, P. E. (1997). Measuring comparability of standards between subjects: Why our statistical techniques do not make the grade. *British Educational Research Journal, 23,* 433–449.

Orlando, M., & Thissen, D. (2000). New item fit indices for dichotomous item response theory models. *Applied Psychological Measurement, 24,* 50–64.

Rubin, D. B. (1976). Inference and missing data. *Biometrika, 63,* 581–592.

Samejima, F. (1969). Estimation of latent proficiency using a pattern of graded scores. *Psychometrika, Monograph Supplement, No. 17.*

Samejima, F. (1973). Homogeneous case of the continuous response model. *Psychometrika, 38,* 203–219.

Smits, N., Mellenbergh, G. J., & Vorst, H. C. M. (2002). Alternative missing data techniques to grade point average: Imputing unavailable grades. *Journal of Educational Measurement, 39,* 187–206.

Takane, Y., & de Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika, 52,* 393–408.

te Marvelde, J. M., Glas, C. A. W., Van Landeghem, G., & Van Damme, J. (2006). Application of multidimensional IRT models to longitudinal data. *Educational and Psychological Measurement, 66,* 5–34.

Thissen, D., Chen, W.-H., & Bock, R. D. (2002). *MULTILOG* [Computer Program and Manual]. Chicago: Scientific Software.

Verhelst, N. D., Glas, C. A. W., & Verstralen, H. H. F. M. (1995). *OPLM* [Computer Program and Manual]. Arnhem: Cito, the National Institute for Educational Measurement, the Netherlands.

Verhelst, N. D., & Verstralen, H. H. F. M. (1993). A stochastic unfolding model derived from the partial credit model. *Kwantitatieve Methoden, 42,* 73–92.

Willms, J. D. (1992). *Monitoring school performance: A guide for educators*. Washington/London: Falmer Press.

Wood, R., Wilson, D. T., Gibbons, R. D., Schilling, S. G., Muraki, E., & Bock, R. D. (2002). *TESTFACT: Test scoring, item statistics, and item factor analysis* [Computer Program and Manual]. Chicago: Scientific Software.

Wu, M. L., Adams, R. J., & Wilson, M. R. (1997). *ConQuest: Generalized Item Response Modeling Software* [Computer Program and Manual]. Camberwell, Victoria: Australian Council for Educational Research.

Young, J.W. (1990). Adjusting the cumulative GPA using item response theory. *Journal of Educational Measurement, 27,* 175–186.

Young, J.W. (1991). Gender bias in predicting college academic performance: A new approach using item response theory. *Journal of Educational Measurement, 28,* 37–47.

Authors

OKSANA B. KOROBKO is a doctoral candidate, Department of Research Methodology at the University of Twente, P.O. Box 217, 7500AE, Enschede, the Netherlands; o.b.korobko@gw.utwente.nl. Her main interests include statistics and psychometrics.

CEES A.W. GLAS is a Professor, Department of Research Methodology at the University of Twente, P.O. Box 217, 7500AE, Enschede, the Netherlands; c.a.w.glas@gw.utwente.nl. His main interests include statistics and psychometrics.

ROEL J. BOSKER is a Professor, Department of Education at the University of Groningen, Grote Rozenstraat 38, 9712 TJ Groningen, the Netherlands; r.j.bosker@rug.nl. His main interests include school effectiveness research.

JOHAN W. LUYTEN is an Assistant Professor, Department of Educational Organization and Management at the University of Twente, P.O. Box 217, 7500AE, Enschede, the Netherlands; j.w.luyten@gw.utwente.nl. His main interests include school effectiveness research.