Statistical Science
1997, Vol. 12, No. 4, 251–278

An Alternative to Traditional GPA for Evaluating Student Performance

Valen E. Johnson

Abstract. In response to the growing problem of grade inflation in American undergraduate institutions, alternatives to GPA and GPA-based student assessment are discussed. One alternative summary, based on a Bayesian latent trait formulation, eliminates many of the inequities associated with GPA-based measures and has been proposed as a replacement for GPA-based class ranks at Duke University.

1. BACKGROUND

Grade point average, or GPA, is the most widely used summary of undergraduate student performance in our educational system. Unfortunately, combining student grades through simple averaging schemes to obtain GPA's results in systematic biases against students enrolled in more rigorous curricula and has important consequences in student course selection. It creates perverse incentives for faculty to inflate grades and lower standards, and it rewards students for selecting less challenging courses and majors (Larkey and Caulkin, 1992).

To understand the problems caused by the use of GPA, consider the data illustrated in Figure 1. This figure depicts boxplots of classroom mean grades for all undergraduate classes with enrollments of 20 or more students offered at Duke University between the fall semester of 1989 and spring semester of 1994 for the 12 departments previously examined in Goldman, Schmidt, Hewitt and Fisher (1974). Two aspects of the grade assignment process are clear from these plots. First, there is substantial variation in the median grades assigned in different departments. Second, there are even greater differences in the grading patterns between instructors within the same departments. As a consequence of these differences, students taking a majority of their classes in, say, Department 1 are likely to finish college with higher GPA's than students who take a majority of their classes in Department 2. In fact, the comparative advantage for Department 1 students may be even greater than indicated in Figure 1, because there is evidence that departments with high-ability students tend to grade more stringently than those with lower-ability students (Goldman and Widawski, 1976).

What portion of the grade differences depicted in Figure 1 can be explained by variations in student achievement levels? Figure 2 is a scatterplot of the mean grade assigned in each Duke department (more precisely, department registration code) against the mean student achievement index for students taking courses in that department. Student achievement indices are described in Section 2, but for present purposes may be considered as an adjusted GPA, adjusted for grading patterns of instructors. As predicted by Goldman and Widawski, this figure suggests that the correlation between mean student achievement and mean grade assigned within departments is quite low or even *negative*.

Such differences in grading patterns have grave implications for our educational system. Besides the obvious inequities inflicted upon students enrolled in "hard" majors, differences in grade distributions result in a substantial reduction in the number of courses taken by students in subjects like mathematics and the natural sciences, as well as other challenging upper-level undergraduate courses. In fact, Larkey and Caulkin (1992) estimated that several hundred thousand fewer mathematics and natural sciences courses may be taken each year in the United States as a direct result of differential grading policies.

Of course, from a student's perspective, avoiding courses in which instructors grade severely is entirely sensible. For those students whose primary

Valen E. Johnson is Associate Professor, Institute of Statistics and Decision Sciences, Duke University, Box 90251, Durham, North Carolina 27708-0251 (e-mail: valen@isds.duke.edu).

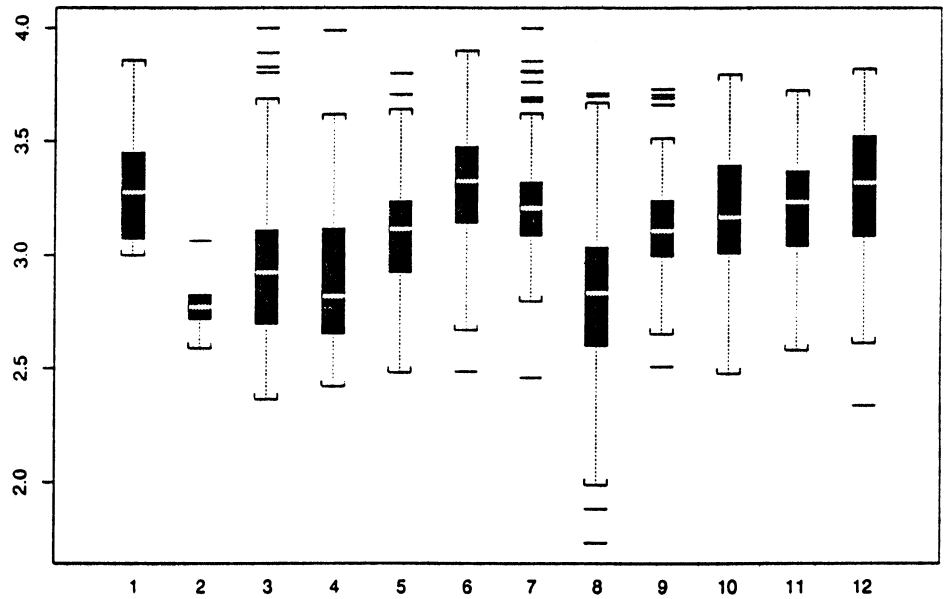

FIG. 1. *Boxplots of mean classroom grades assigned in classes offered by* 12 *Duke University departments. The* 12 *departments coincide with the* 12 *departments studied in Goldman et al.* (1974): *cultural anthropology, biochemistry, biology, chemistry, economics, engineering, history, mathematics, political science, psychology, sociology and Spanish.*

objective is to gain admittance to medical school or law school, or to land a lucrative position on Wall Street, it may be irrational to take any but the required courses in hard departments, or from instructors who grade severely. For such students, "grade shopping" may represent an optimal career strategy.

For similar reasons, inflating grades is a reasonable strategy for faculty members, especially junior faculty. By assigning higher than average grades, course enrollments increase, student complaints are minimized and students spend less time during office hours negotiating for higher marks. Additionally, salary increases, promotions and tenure

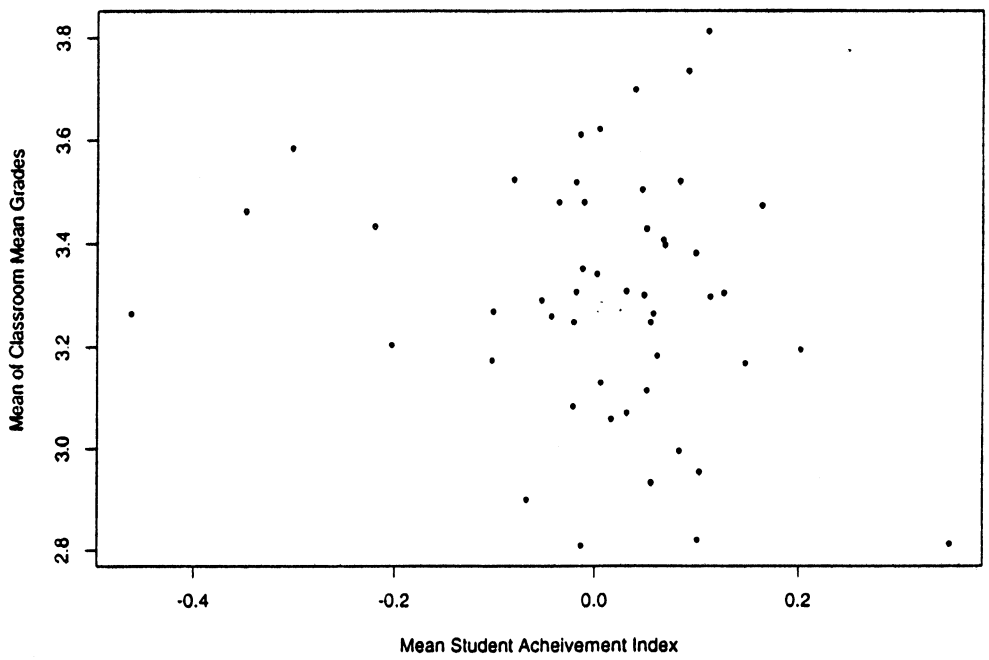

FIG. 2. *Scatterplot of the mean of the mean classroom grades assigned by Duke course registration code versus the mean student achievement index of students receiving these grades. Registration codes in which fewer than* 100 *grades were assigned were not plotted.*

decisions are often tied to student course evaluations, which in turn are positively correlated with assigned grades. Practically speaking, there seems to be little faculty incentive for not inflating grades.

What then can be done to remove the disincentives for learning caused by differential grade assignment? A number of alternate measures for student performance have been proposed in the educational literature. In this article, I briefly review the more prominent of these, and propose a Bayesian model for student assessment that incorporates the essential ideas from each. This model represents a variation of the Bayesian ordinal data models proposed in Albert and Chib (1993), Johnson (1996) and Cowles, Carlin and Connett (1996) and may thus be regarded as a Bayesian extension of the graded response models described in, for example, Samejima (1969) and Young (1990). Graded response models in turn find their roots in item response theory (e.g., Lord and Novick, 1968), which is now omnipresent in the quantitative psychology literature (see, e.g., Fischer and Molenaar, 1995, and van der Linden and Hambleton, 1997, for extensive reviews of research in this class of models). By adopting this alternate measure of student performance, penalties imposed on students for taking challenging courses in their major fields of study can be eliminated, and incentives for learning can be reintroduced into the system.

The essential idea that motivates this adjustment method is that the relative rankings of students within classes, rather than absolute grades, should be used to evaluate student performance. Under the proposed model for ranking student performance, an instructor who assigns A's to all students in a class provides exactly the same information as an instructor who assigns C's to the same students when enrolled in another class. Because the model adjusts for instructor differences in grading policies, much of the subjectivity involved in the interpretation of assigned grades is eliminated. Furthermore, the model automatically adjusts for the achievement levels of students within a class, by comparing the relative ranking of students as they mix across classes. For example, though Students A and C may not take any classes together, information that Student A did better than Student B in one class, and that Student B did better than Student C in another, provides information about the relative achievement levels of Students A and C.

Before describing this new proposal for student assessment, it is useful first to review the earlier proposals briefly. More complete reviews can be found in Linn (1966) and Young (1993).

1.1 Pairwise-Comparisons Methods

Goldman and Widawski (1976) proposed a grade adjustment method based on pairwise comparisons of grades obtained by the same students across multiple departments. In their method, the difference in a student's grades for classes taken in different departments provides information about the relative grading standards between the departments. Goldman and Widawski averaged all such differences obtained from the transcripts of 475 University of California at Riverside (UCR) students to obtain grade adjustment factors for 17 academic fields. Based on this analysis, Goldman and Widawski concluded that there were systematic differences in grading patterns across academic departments at UCR and that departments with high-ability students tended to grade more stringently than fields with less able students.

Goldman and Widawski's analysis was extended by Strenta and Elliott (1987) and Elliott and Strenta (1988) in studies of Dartmouth College undergraduates. In Strenta and Elliot, pairwise comparisons of departments were restricted to introductory courses, and external measures of student abilities (SAT and high school GPA) were used for validation. The results corroborate the earlier conclusions of Goldman and Widawski and confirm a stable trend in differential department grading standards over a 10-year period and between public and private institutions. In their later article, Elliott and Strenta incorporated both within and between department course comparisons, and also estimated grade adjustments for a larger number of departments. Once again, resulting indices produced adjusted GPA measures that correlated more strongly with both SAT and high school GPA than did standard GPA measures.

1.2 Graded Response Models

In his 1989 doctoral thesis, Young adapted a model derived from item response theory (IRT; e.g., Lord and Novick, 1968) called the graded response model (GRM; Samejima, 1969) for application to undergraduate grade data. In this model, it is assumed that there are a total of K grades which can be assigned to students and that these grades are numbered and ordered from 1 to K. The grade assigned to student i in class j is denoted by Y_{ij}, while the underlying variable representing the ith student's ability is denoted by β_i. In the terminology of IRT, let η_j denote the discrimination parameter of the jth class grade, and let ζ_{jk} denote the upper grade-cutoff for grade k in class j. With this notation, the basic assumption of the GRM is

that

$$\Pr[Y_{ij} \le k] = \pi_{ijk} = \frac{\exp[\eta_j(\beta_i - \zeta_{jk})]}{1 + \exp[\eta_j(\beta_i - \zeta_{jk})]}. \quad (1)$$

An important feature of the GRM is that it explicitly parameterizes the grade-cutoffs for each class. In theory, this allows the GRM to account for variations in instructor grading patterns.

Young applied this model to a cohort of Stanford undergraduate grades and found that estimates of student abilities obtained using this model correlated better with external measures of student abilities than did raw GPA (Young, 1990, 1993). For example, the multiple correlation of student abilities obtained from the GRM model with verbal and mathematics SAT scores and high school GPA was higher than it was for raw GPA.

From a technical standpoint, it is clear from (1) that the parameters $\{\eta_j\}$, $\{\beta_i\}$ and $\{\zeta_{jk}\}$ are not identifiable. To see this, note that η_j, β_i and ζ_{jk} may be replaced by $s\eta_j$, $(\beta_i + c)/s$ and $(\zeta_{jk} + c)/s$ for arbitrary constants c and s without affecting the predicted value of π_{ijk}.

In order to make model parameters identifiable, disciples of maximum likelihood estimation typically assume both a probabilistic constraint (prior density) on student abilities, and fixed constraints on two of the extreme grade-cutoffs (e.g., Young, 1990; Muraki, 1990; see also Bradlow, 1994, Bradlow and Zaslavsky, 1996 and Nandram and Chen, 1996, for related discussion on Bayesian models employing fixed grade-cutoffs). Unfortunately, fixing grade-cutoffs can have deleterious effects when the model is used to produce student rankings. To see this, suppose that the lower cutoff for an A+ is fixed at a constant value of, say, 3. Then the classroom performance of students receiving an A+ in *any* class must be estimated to have a value in excess of 3. But when all grades assigned to students in a given course are A+'s, this assumption results in an inflation of the class rankings of students in the class relative to all other students, even if the average achievement of students taking the class is below average (assuming that average student ability is centered at 0). This effect is discussed in more detail in Section 2, where a Bayesian variation of a model similar to the GRM is proposed.

1.3 Regression Models

More recently, Larkey and colleagues at Carnegie Mellon University investigated linear techniques for adjusting student GPA's to account for the difficulty of courses taken (Caulkin, Larkey and Wei, 1996; Larkey and Caulkin, 1992; Larkey, 1991; see also Young, 1992). In the simplest and perhaps most useful version of their approach, an additive adjustment is made to each student's GPA based on estimates of the difficulty of the student's curriculum. The difficulty of courses is estimated from a linear regression of student grades on "true" student GPA and course difficulty parameters. If Y_{ij} again denotes the grade of the ith student in the jth class, and g_i and c_j denote the ith student's true GPA and the difficulty of the jth class, respectively, then the additive model used to estimate these adjustment factors takes the form

$$Y_{ij} = g_i - c_j + e_{ij}. \quad (2)$$

In (2), e_{ij} represents a mean zero, normally distributed error term. Numerical procedures for estimating model parameters are described in Caulkin, Larkey and Wei (1996).

Caulkin, Larkey and Wei report that predictions obtained using the additive adjustment model produced estimates of student performance that correlated more highly with high school GPA and SAT scores than did estimates obtained using the GRM. The performance of the additive adjustment model on a cohort of Duke University students is examined further in Section 4.2.

2. A BAYESIAN MODEL FOR GRADE DATA

Each of the grade adjustment methods described above attempts to account simultaneously for the two critical factors that affect grade assignment: the achievement levels of students within a class, and instructor-specific grade cutoffs relative to these perceived achievement levels. To account formally for both of these factors, I propose a synthesis of the models described in Albert and Chib (1993), Johnson (1996) and Cowles, Carlin and Connett (1996) for ordinal and multirater ordinal data.

The proposed model begins with the assumption that an instructor assigns grades by first ordering perceived student performance from best to worst, possibly with ties for groups of students who performed at approximately the same level. This ranking is assumed to be based on an achievement index that is implicitly defined by the instructor. The particular definition of achievement for each class is arbitrary.

After ordering students according to their estimated classroom achievement, instructors next group students into grade categories by fixing grade-cutoffs between the estimated achievement levels of students in the class. In some instances, instructors base grade-cutoffs on a predetermined notion of the knowledge levels required for each

grade. In others, grade-cutoffs are determined using a "curve," whereby instructors attempt to assign a certain proportion of students to each grade level. Regardless of how grades and the corresponding grade-cutoffs are assigned, it is important to emphasize that the manner in which an instructor determines grade-cutoffs is not critical in the calculation of the achievement indices in the model described below. Only the relative ordering of students within classes, as determined by assigned grades, provides information about student achievement.

The following variables are used to model this mechanism for grade generation:

1. The variable Y_{ij} denotes the grade assigned to the ith student in the jth class. For notational convenience, the grades used at Duke were coded so that F = 1, D− = 2, D = 3, ..., A+ = 13. In general, there are K possible grades, ordered from 1 to K.
2. The variable X_i represents the mean classroom achievement of the ith student, *in classes selected by student i*. This variable is called the *achievement index* (AI) of student i. It should be noted that this definition of the "achievement index" differs from the standard definition of a latent trait (e.g., Lord and Novick, 1968) in that the achievement index is defined conditionally for those classes selected by a student. In contrast, the usual definition of a latent trait is not applicable in this setting due to the confounding effect of student course selection (e.g., Wang, Wainer and Thissen, 1995).
3. Grade-cutoffs for class j are denoted by γ_0^j, $\gamma_1^j, \ldots, \gamma_K^j$. The upper cutoff for an F in class 3 is γ_1^3, for a D– it is γ_2^3 and so on up to γ_{12}^3, which is the upper cutoff for an A. The upper cutoff for an A+, γ_{13}^j, is ∞, and the lower cutoff for an F, γ_0^j, is $-\infty$.
4. Random variation in the performance of student i in class j is denoted by ε_{ij}. This term accounts for the fact that student achievement varies from class to class and that instructor assessment of student achievement is also subject to error. It is assumed that the distribution of each ε_{ij} is Gaussian with mean 0 and variance σ_{ij}^2. In addition, the variation of observed student achievement in class j is assumed to depend only on the instructor of class j, denoted by $t(j)$. That is, it is assumed that $\sigma_{ij}^2 = \sigma_{t(j)}^2$, independently of i.

With these variable definitions, the model for grade generation may be summarized as follows. Student i gets a grade of $Y_{ij} = k$ in class j if and only if

$$\gamma_{k-1}^j < X_i + \varepsilon_{ij} \le \gamma_k^j. \tag{3}$$

It follows from (3) that the probability that student i receives a grade of k in class j is equal to the area under the normal curve within category k. Letting $\Phi(\cdot)$ denote the cumulative standard normal distribution function (and $\phi(\cdot)$ the standard normal density), this area may be expressed

$$\Phi\left(\frac{\gamma_k^j - X_i}{\sigma_{t(j)}}\right) - \Phi\left(\frac{\gamma_{k-1}^j - X_i}{\sigma_{t(j)}}\right). \tag{4}$$

If we assume that the grades received by students are independent given model parameters, the likelihood function for a cohort of grade data is the product over all probabilities of the form (4), or

$$\prod_i \prod_{j \in \mathscr{C}_i} \left[\Phi\left(\frac{\gamma_{Y_{ij}}^j - X_i}{\sigma_{t(j)}}\right) - \Phi\left(\frac{\gamma_{Y_{ij}-1}^j - X_i}{\sigma_{t(j)}}\right)\right]. \tag{5}$$

The first product in (5) extends over all students, while the second extends over the set of all classes j taken by student i, denoted here by $\mathscr{C}_i$.

As in the GRM, the addition of a constant to all category cutoffs and achievement indices does not affect the likelihood function. Similarly, the likelihood is unchanged if all quantities are divided by a scalar constant. This is a consequence of the fact that grades are ordinal and possess no natural scale. One possibility for dealing with this dilemma is simply to fix the values of two of the category cutoffs: this was the approach taken by Young (1990) in the GRM model. However, this approach leads to the inconsistencies in the model specification detailed in Section 1.2.

A more natural way to establish an underlying scale of measurement is to specify the marginal distribution of the achievement indices, category cutoffs and error terms. From within the Bayesian paradigm, this is accomplished through the introduction of prior distributions on the unknown quantities of interest.

The prior distributions used in this model are

$$X_i \sim N(0, 1), \tag{6}$$

$$\varepsilon_{ij} \sim N(0, \sigma_{t(j)}^2) \tag{7}$$

and

$$\sigma_{t(j)}^2 \sim \mathrm{IG}(\alpha, \lambda) \propto (\sigma_{t(j)}^2)^{-(\alpha+1)} \exp\left(-\frac{\lambda}{\sigma_{t(j)}^2}\right). \tag{8}$$

Alternative prior densities are considered in Section 4.2.2. In the baseline model, $\alpha = 1.5$ and $\lambda = 1.5$. Throughout, $N(\mu, \tau^2)$ represents a normal distribution with mean μ and variance τ^2, and $\mathrm{IG}(\alpha, \lambda)$

denotes an inverse gamma distribution with shape α and scale λ. Recall that the subscript $t(j)$ refers to the teacher of class j.

The prior density on the achievement indices (6) is standard in IRT literature and simply serves to fix the scale of measurement.

Equation (7) states that the random error associated with the combination of interclass variation in student performance and instructor assessment error follows a mean-zero normal distribution on the scale of the achievement indices. It is further assumed that the variance of the error depends only on the instructor of class j. This variance is denoted by $\sigma^2_{t(j)}$.

Assumption (8) describes the marginal distribution on the variance of the random errors. The particular parameters selected for the inverse gamma distribution lead to a relatively vague prior on the instructor variances and were chosen to satisfy several criteria. Among these, the positive value of λ eliminates an irregularity in the posterior distribution that occurs when all category cutoffs, achievement indices and instructor variances assume values close to 0. For $\alpha = \lambda = 1.5$, the priors on each instructor variance have mean 3 and mode 0.6, which seems consistent with the assumption that the marginal distribution on student achievement indices has variance 1. The parameter estimates produced using these values of α and λ are similar to those obtained using both an empirical Bayes approach for estimating α and λ, and those obtained by taking a uniform prior on α and λ and treating these parameters as random quantities as well. Results from the empirical Bayes and fully Bayesian approach for estimating α and λ are summarized in Section 4.2.2. From a technical standpoint, the inverse gamma structure facilitates sampling from the posterior distribution on the achievement indices using the data augmentation scheme described in Section 3.1.

Defining appropriate prior distributions for the grade-cutoffs requires more careful consideration of the mechanisms underlying grade assignment. For example, if a locally uniform prior (subject to the ordering constraint $\gamma^j_k \le \gamma^j_{k+1}$) is taken on the grade-cutoffs, the joint posterior distribution on the achievement indices concentrates its mass above its prior mean of 0. This occurs because unobserved grade categories, which usually correspond to lower grades, are assigned nonnegligible probability in the posterior. In other words, a uniform prior for the grade-cutoffs would assign equal prior probability to all grades, while in practice, grades below C are relatively rare. Thus, if a uniform (or other reference) prior were employed for the grade-cutoffs, the posterior probability assigned to below-C grades, particularly in small classes, would be nonnegligible, forcing the posterior distribution on the achievement indices to become highly skewed. As a consequence, the MAP estimate would assign zero probability to unobserved grades, while samples from the posterior would assign positive probability to the same grades. The nonnegligible mass assigned to low grades would then force the posterior mean of the achievement indices upward toward the inflated grade-cutoffs that correspond to the actual grades assigned.

In addition to the effect of unobserved grades on the shape of the posterior, reference priors on grade-cutoffs lead to posterior distributions that are not invariant to shifts in class grades. That is, shifting the grades of all students within a class down by one letter grade (assuming no grades of D$-$ or below) does not change the ordering of students within a class and so should not affect the rankings of students. Yet the use of uniform or other vaguely specified priors on the grade-cutoffs lead to different estimates of student achievement indices after such shifts.

Priors which lead to posteriors that concentrate negligible mass in unobserved grade categories are therefore needed. Such priors can be specified by introducing binary random variables $(\iota^j_1, \ldots, \iota^j_K) = \boldsymbol{\iota}_j$ to indicate which grade-cutoffs correspond to unobserved grades. By placing high prior probability on the occurrence of unobserved grade-categories, the posterior probability assigned to unobserved categories can be made arbitrarily small. One prior density that satisfies these criteria can be specified by first assigning probability 1 to the event $\gamma^j_{k-1} = \gamma^j_k$ whenever $\iota^j_k = 0$, and then taking the prior density on $\boldsymbol{\iota}_j$ to be

$$\Pr(\boldsymbol{\iota}_j) = \begin{cases} \dfrac{\varepsilon^{S_j}}{1-(1-\varepsilon)^K}, & \text{if } S_j \ge 1, \\ 0, & \text{if } S_j = 0, \end{cases} \tag{9}$$

$$S_j = \sum_{k=1}^{K} \iota^j_k,$$

where $\varepsilon \ll 1$.

It follows that the prior conditional distribution for a component of $\boldsymbol{\iota}_j$, given all other components is

$$\Pr\Bigg(\iota^j_i = 1 \Bigg| \sum_{\substack{k\ne i \\ 1\le k<K}} \iota^j_k \ge 1\Bigg) = \frac{\varepsilon}{1+\varepsilon}, \tag{10}$$

$$\Pr\Bigg(\iota^j_i = 0 \Bigg| \sum_{\substack{k\ne i \\ 1\le k<K}} \iota^j_k = 0\Bigg) = 0.$$

The constraint that $\sum \iota_k^j \geq 1$ insures that $-\infty = \gamma_0^j \neq \gamma_K^j = \infty$, or in other words, that the lower cutoff for an F cannot equal the upper cutoff for an A+. Of course, whenever a grade of k is observed in class j, the posterior probability that $\iota_k^j = 1$ is 1.

Given $\boldsymbol{\iota}_j$, a non-informative (Jeffrey's) prior is assumed for the unique elements of the grade cutoffs in the jth class. Because the achievement indices are assumed to have a $N(0, 1)$ distribution, the Jeffrey's prior on the probability that a student receives a given grade is transformed to this scale, resulting in a prior density of the form

$$p(\boldsymbol{\gamma}^j) \propto \Phi(\gamma_{I_{\min}})^{-1/2}\left\{\prod_{\substack{i_k \in I \\ i_k > I_{\min}}} \left[\Phi(\gamma_{i_k}^j) - \Phi(\gamma_{i_{k-1}}^j)\right]\right\}^{-1/2} \cdot \left\{\prod_{\substack{i_k \in I \\ i_k < I_{\max}}} \phi(\gamma_{i_k}^j)\right\}. \tag{11}$$

Here $I = \{k\colon \iota_k = 1\}$, i_k denotes the $(k+1)$st largest element of I, and $I_{\min}$ and $I_{\max}$ refer to the smallest and largest elements of I.

3. PARAMETER ESTIMATION

In practice, numerical strategies for both sampling from the posterior distribution and for obtaining point estimates of the achievement indices are needed. Sampling strategies are critical for assessing model fit and comparing models, while a method for rapidly obtaining point estimates of the achievement indices is important for implementation by a registrar's office.

3.1 Posterior Simulation

Sampling from the posterior distribution of model parameters can be accomplished using Markov chain Monte Carlo (MCMC) techniques (see, e.g., Gilks, Richardson and Spiegelhalter, 1996). The particular sampler used here is based on a Metropolis–Hastings (MH) algorithm that utilizes the Bayesian data augmentation (BDA) scheme proposed by Tanner and Wong (1987). This algorithm is similar to the one employed by Albert and Chib (1993) for probit models and the multirater ordinal data models described in Johnson (1996) and Cowles, Carlin and Connett (1996).

The BDA/MH scheme can be implemented by introducing variables that represent the unobserved classroom achievement of each student in each class. To this end, define the classroom achievement for the ith student in the jth class to be $x_{ij} = x_i + \varepsilon_{ij}$. It follows that the likelihood function based on the parameter space augmented with the vector $\{x_{ij}\}$ may be written

$$\prod_i \prod_{j \in \mathscr{C}_i} \phi\left(\frac{x_{ij} - x_i}{\sigma_{t(j)}}\right) \mathrm{Ind}\left(\gamma_{Y_{ij}-1}^j < x_{ij} \leq \gamma_{Y_{ij}}^j\right) \tag{12}$$

In (12), Ind$(\cdot)$ represents the indicator function taking a value of 1 if the stated condition is true, and 0 otherwise. Note that integration of (12) over the variables $\{x_{ij}\}$ leads to (5).

Based on the augmented parameter space, an MCMC algorithm can be defined in a straightforward way using the steps detailed in Johnson (1996) and Cowles, Carlin and Connett (1996).

3.2 Posterior Optimization

As in the case of related GRM's, optimization over the parameter space requires judicious choice of initial values. Although the prior densities (6)–(11) insure identifiability, the form of the non-informative priors on the grade-cutoff vectors and the instructor variance parameters results in nonconcavity of the log-posterior density.

Despite this nonconcavity, MAP estimates for student achievement indices can be reliably obtained using a variation of the ICM algorithm (Besag, 1986), provided that reasonable starting values are used for parameter initialization. As in the standard ICM algorithm, the optimization strategy used here evolves by sequentially maximizing the conditional distribution of each component of the parameter vector, given current estimates of all other components. Collapsed grade-cutoffs are updated simultaneously to avoid premature "freezing."

Convergence of the algorithm is normally obtained within 200 iterations when applied to data sets with 1,500 students each taking approximately 35 courses from 2,000 instructors, and requires 4 hours on a midpriced Unix workstation.

4. CASE STUDIES

In this section, the Bayesian model for student achievement is applied to two data sets. The first is a stylized example borrowed from Larkey and Caulkin (1992) in which the Bayesian achievement indices are shown to produce student rankings that are exactly opposite those obtained using GPA. Although this particular example is extreme, the irregularities in its grade assignments are not atypical of actual college transcripts. The example also clarifies the underlying problems associated with GPA-based student assessment.

The second example involves an analysis of grades received by a recent class of Duke University undergraduates. The analysis begins with a

TABLE 1
Larkey–Caulkin example

	Student I	Student II	Student III	Student IV	Class GPA
Class 1	B+			B−	3.0
Class 2	C+		C		2.15
Class 3			A	B+	3.65
Class 4	C−	D			1.35
Class 5		A		A−	3.85
Class 6	B+			B	3.15
Class 7		B+	B		3.15
Class 8	B+	B	B−	C+	2.83
Class 9		B	B−		2.85
GPA	2.78	2.86	2.88	3.0	

comparison of AI-derived student class rank and class ranks obtained using unadjusted GPA and additively adjusted GPA. Two performance measures are used in this comparison: one is based on the multiple correlation of each index with external measures of student performance; the second on the predictive success of the indices in ordering student performance within individual classes.

Following this comparison, the statistical properties of the Bayesian model are investigated in greater detail. The primary questions addressed in the subsequent analysis concern (1) the adequacy of the Bayesian model in representing observed variation in assigned grades and (2) the sensitivity of student rankings to underlying model assumptions. Clearly, satisfactory resolution of these issues is a prerequisite for adopting this model as a replacement for traditional GPA.

The first question is explored through an examination of latent residuals. The second requires more careful attention and focuses on two issues: the validity of representing student achievement with a univariate quantity, and the sensitivity of student rankings to the particular priors employed for the achievement indices, grade-cutoffs and instructor variance parameters. The former issue is explored by comparing the one-component model to a two-component model, the latter by modifying prior assumptions and observing concomitant changes to the posterior. The conclusions of this analysis are that estimated student ranks are relatively unaffected by moderate perturbations of second-stage model assumptions.

4.1 Larkey–Caulkin Data

The data in Table 1 represents a slight modification of data originally presented by Larkey and Caulkin (1992). In their version of this example, the letter grades of Table 1 had numerical values on a 100-point scale.

The important feature of this grade data is that all instructors agree that the best ordering of students is I > II > III > IV. Yet, because of differences in instructor grading policies, the observed ranking based on GPA is IV > III > II > I, exactly the *opposite* of the ranking intended by *all* instructors.

To illustrate the properties of the student achievement index, the Bayesian model was also applied to this data. Output from the model is displayed in Table 2. The columns in this table represent (1) class number, (2) grade received, (3) mean grade assigned in the class, (4) estimated classroom achievement index of the student, (5) mean achievement index of all students enrolled, (6) mean GPA of all students in the class and (7) estimated category cutoffs for the grade received. Student GPA's appear at the bottom of the columns labeled "Grade," and student achievement indices are listed at the bottom of the column labeled "Estimated achievement." The values of $\pm\infty$ were coded as ± 9.99. All quantities in the table were normalized so that the posterior mean and variance of the achievement indices were 0 and 1, respectively.

The salient feature of Table 2 is that the rank of students based on the achievement index is correct: I is ranked first, II second, III third and IV fourth.

To gain further insight into the meaning of model parameters, consider the grade of D received by Student II in Class 4. The estimate of this student's achievement in this class was -0.08, which on the probability scale corresponds to the 47th percentile. The mean achievement index for all students taking this class was 0.79, indicating that better than average students were enrolled in the course, and the mean grade assigned in Class 4 was 1.35. Because of the higher than average achievement level of students enrolled in Class 4, and the lower than average grade assigned, the grade-cutoffs for a D were estimated to be between $-\infty$ and 0.59. By comparison, the C cutoffs in Class 2 were $(-\infty, 0.24)$.

TABLE 2
Analysis of Larkey–Caulkin example

Course	Grade	Mean grade	Estimated achievement	Mean achievement	Mean GPA	Grade cutoffs
			Student I: GPA-based rank, 4; achievement index rank, 1			
CLS 001	B+	3.00	1.25	−0.00	2.89	(−0.00, 9.99)
CLS 002	C+	2.15	1.32	0.36	2.83	(0.24, 9.99)
CLS 004	C−	1.35	1.45	0.79	2.82	(0.59, 9.99)
CLS 006	B+	3.15	1.25	−0.00	2.89	(−0.00, 9.99)
CLS 008	B+	2.83	1.51	0.00	2.88	(0.71, 9.99)
	2.78	2.50	1.15	0.23	2.86	
			Student II: GPA-based rank, 3 achievement index rank, 2			
CLS 004	D	1.35	−0.08	0.79	2.82	(−9.99, 0.59)
CLS 005	A	3.85	0.69	−0.36	2.93	(−0.24, 9.99)
CLS 007	B+	3.15	0.80	0.00	2.87	(−0.00, 9.99)
CLS 008	B	2.83	0.36	0.00	2.88	(0.00, 0.71)
CLS 009	B	2.85	0.80	0.00	2.87	(−0.00, 9.99)
	2.86	2.81	0.43	0.09	2.87	
			Student III: GPA-based rank, 2; achievement index rank, 3			
CLS 002	C	2.15	−0.69	0.36	2.83	(−9.99, 0.24)
CLS 003	A	3.65	0.08	−0.79	2.94	(−0.59, 9.99)
CLS 007	B	3.15	−0.80	0.00	2.87	(−9.99, −0.00)
CLS 008	B−	2.83	−0.36	0.00	2.88	(−0.71, 0.00)
CLS 009	B−	2.85	−0.80	0.00	2.87	(−9.99, −0.00)
	2.88	2.92	−0.43	−0.09	2.88	
			Student IV: GPA-based rank, 1; achievement index rank, 4			
CLS 001	B−	3.00	−1.25	−0.00	2.89	(−9.99, −0.00)
CLS 003	B+	3.65	−1.45	−0.79	2.94	(−9.99, −0.59)
CLS 005	A−	3.85	−1.32	−0.36	2.93	(−9.99, −0.24)
CLS 006	B	3.15	−1.25	−0.00	2.89	(−9.99, −0.00)
CLS 008	C+	2.83	−1.51	0.00	2.88	(−9.99, −0.71)
	3.00	3.30	−1.15	−0.23	2.91	

4.2 A Class of Duke University Undergraduates

To illustrate the performance of the Bayesian achievement indices for actual college transcripts, and to compare its performance to raw GPA and additively adjusted GPA, all models were applied to the grades of a recent class of Duke University undergraduates. Selected transcript summaries from the Bayesian model are displayed in the Appendix (to safeguard the privacy of students and instructors, the terms, years and final course digit-designators of all classes were omitted). Transcripts in which the student class rank based on GPA differed sharply from achievement-index-based rank were chosen for display. Approximately 1,400 students were ranked. Perusal of these transcripts is left to the reader.

Figure 3 is a scatterplot of GPA rank against achievement index rank for this cohort. As illustrated, the correlation between the two measures of student performance is relatively high, and for many students the differences in class rank obtained from the two indices are not large. However, for other students, the rank percentiles may differ by as much as 40%, and so the effect of this reform can be quite substantial.

Given the disparities between GPA rank and achievement-based rank, an obvious question becomes "Which rank better represents student performance?" Two criteria were used to address this question. The first, which appears to be the most commonly used statistic for assessing alternative measures of undergraduate student performance (e.g., Elliot and Strenta, 1988; Young, 1990; Larkey and Caulkin, 1992; and Caulkin, Larkey and Wei, 1996), was based on the multiple R^2 of the regression of high school GPA, math SAT and verbal SAT scores on each competing measure of student performance. Because the explanatory variables in these regressions are measured independently of

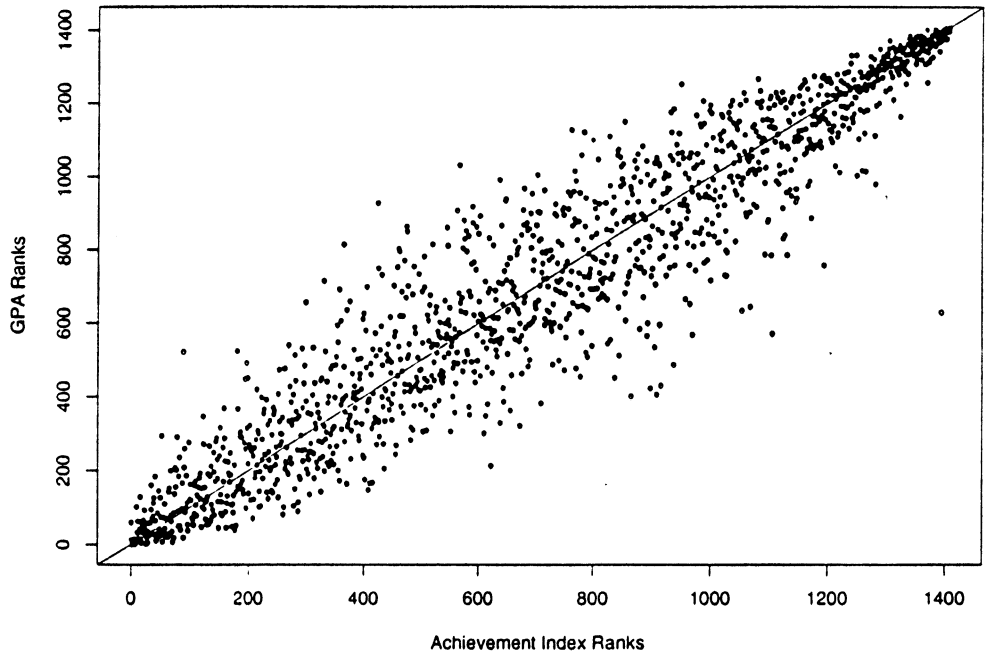

FIG. 3. *Scatterplot of student ranks based on GPA versus student ranks based on achievement index for a recent class of Duke undergraduates.*

college GPA and achievement-based ranks, the resulting R^2 values provide an external measure for model assessment. A potential problem with this measure is that the student selection processes that influence college GPA are also likely to influence high school GPA, and so might favor models linked to raw GPA adjustments.

For this cohort of undergraduate grades, the multiple R^2 for the regression of college GPA on high school GPA and math and verbal SAT scores was 0.252. The regression of achievement indices estimated from the ICM algorithm for the same covariates was 0.346, a substantial increase over the value obtained using raw GPA. For the additive adjustment model, the R^2 value was 0.338. This appears to be in general agreement with the value reported in Larkey and Caulkin (1992) of 0.321 for a selected subset of Carnegie Mellon undergraduates.

A second criterion for comparing performance indices can be based on the power of the indices in predicting the relative performance of students. For example, an effective index should accurately predict the better of two students, or, equivalently, which student is likely to receive the higher grade in a class they take together. Of course, implementing this criterion requires that the better student be known, and no gold standard exists for making this determination. Complicating the issue further is the fact that there is natural variation in the classroom performance of students, so the better student does not always obtain the higher mark.

Fortunately, the predictive accuracy of achievement indices for pairs of students who receive different grades in each of two courses taken together can be accurately assessed. Supposing that an ideal performance index was available, let the probability that the "better" student receives the higher grade in a randomly selected course be denoted by p. If grades in distinct courses are assumed to be independent, then the probability that the better student, according to the ideal performance index, receives the higher grade in both courses is p^2; that the better student receives the lower grade in both courses is $(1-p)^2$. Similarly, the probability that each student gets one of the higher grades is $2p(1-p)$.

Based on this observation, a simple estimate of p can be obtained by equating $2p(1-p)$ to the observed proportion of times students in such matched pairs each received one of the higher marks. That proportion was 0.227 for the Duke cohort, which suggests that an ideal performance index would have an error rate of at least $p = 0.131$. Thus, 0.131 provides an estimate of the baseline error rate for *any* index, and we can compare the prediction errors of other indices to this figure, within this subset of student grades. For the Bayesian achievement indices, additively adjusted GPA and

raw GPA, the corresponding error rates were 0.168, 0.174, and 0.206, respectively. Thus, the error rate attributable to model inadequacy was bounded by 0.037 for the Bayesian achievement indices, while the model-based error for the additively adjusted GPA model was bounded by 0.043. Raw GPA yielded a model-based error of 0.075, twice that incurred using the Bayesian achievement indices. As in the case of the R^2 criteria, the Bayesian indices appear to provide the most accurate estimates of student achievement.

4.2.1 *Residual analyses*

Aside from examining the predictive power of student achievement indices, and their correlation with external measures of student performance, it is also important to assess the adequacy of the model in describing the observed variation in the data: in other words, to assess model fit. Assessing model fit from within the Bayesian framework is often straightforward, and in the present case model fit was evaluated through an examination of latent residuals for unobserved classroom achievement. These residuals were defined for each assigned grade according to the prescription

$$r_{ij} = \frac{x^*_{ij} - x_i^{-j}}{\sigma_{t(j)}}. \tag{13}$$

Here, x_i^{-j} is the estimated posterior mean of the ith student achievement index, excluding the contribution from Y_{ij}. The quantity x^*_{ij} represents a randomly sampled value of the latent classroom achievement of student i in class j, as defined in (12), but based on x_i^{-j} instead of x_i. Numerical techniques for computing the components of x_i^{-j} within the full BDA/MH scheme appear in Gelfand (1996).

The definition of these residuals is similar to the definition of latent residuals for binary regression models presented in Albert and Chib (1995). Given x_i^{-j} and $\sigma^2_{t(j)}$, the distribution of x_{ij} is $N(x_i, \sigma^2_{t(j)})$. Ignoring the dependence in the posterior distribution of σ_j^2 on Y_{ij}, it follows that the conditional distribution of r_{ij} should be approximately $N(0, 1)$.

In order to estimate the quantities x_i^{-j} numerically and to sample values of x^*_{ij} and $\sigma_{t(j)}$, 70,000 iterations of the BDA/MH algorithm were performed. The initial sample of 20,000 iterations was discarded to allow for burn-in of the chain. Values of x_i^{-j} were obtained by averaging sampled values obtained in the last 50,000 iterations, and sampled values for x^*_{ij} and $\sigma_{t(j)}$ were obtained by generating an additional sample of 30,000 iterations and subsampling every 2,000 iterations.

A normal quantile–quantile plot for the last subsampled set of residuals is displayed in Figure 4. Deviations of the data from the model are reflected through departures of the sorted residuals from the indicated line of slope 1 and intercept 0. The lack of departures from the line suggests that the model provides an adequate representation of observed variability in the data. The appearance of this plot

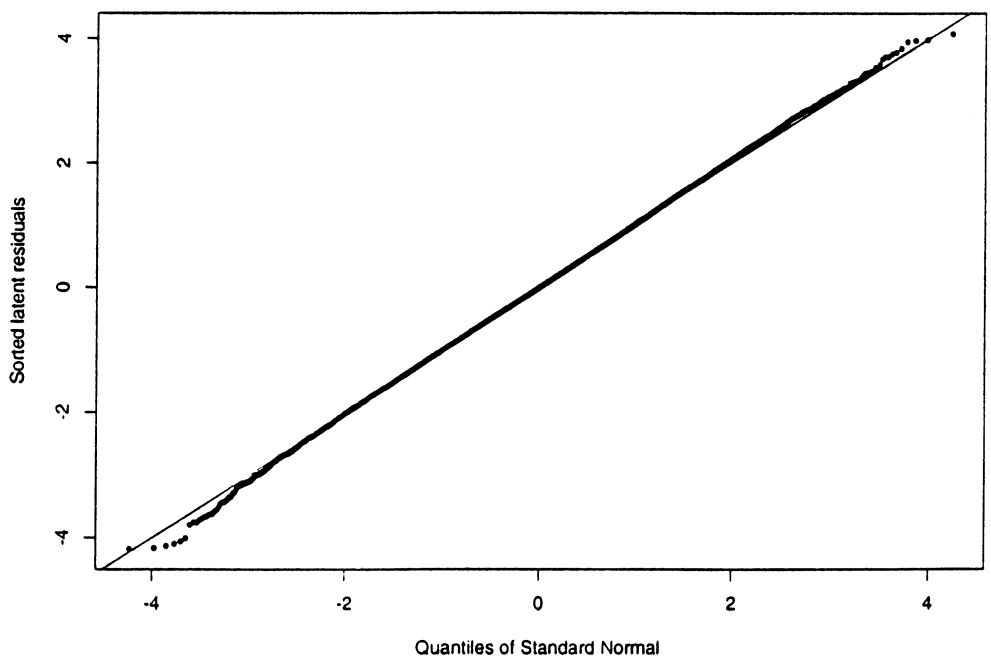

FIG. 4. *Normal scores plot of sampled latent residuals.*

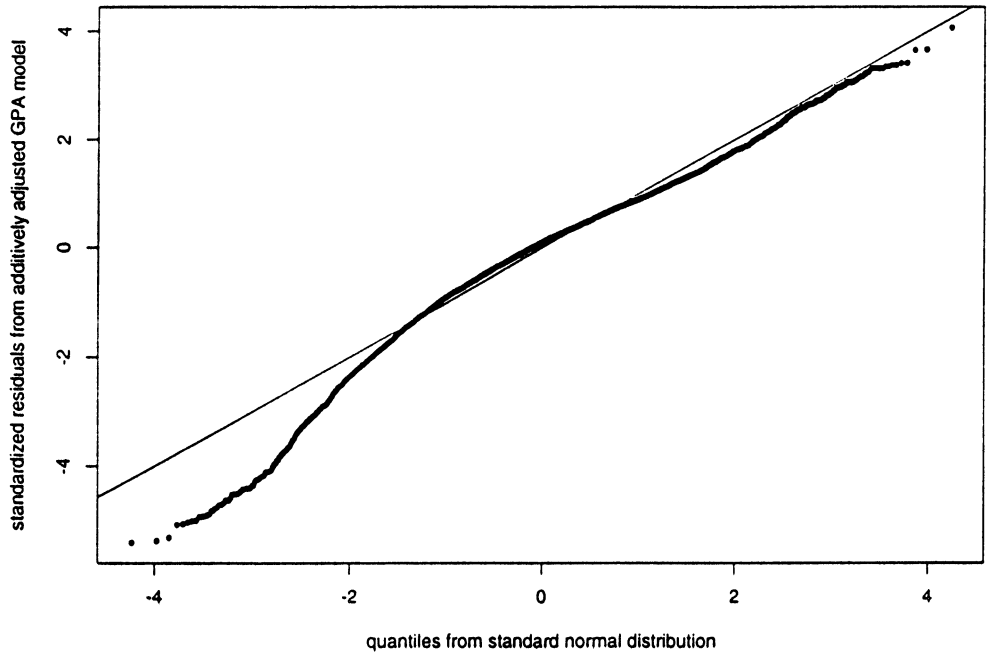

FIG. 5. *Normal scores plot of sampled latent residuals.*

is typical of the appearance of the plots obtained with other residual samples. For comparison, sampling theory residuals from the additive model are displayed in Figure 5.

4.2.2 *Alternative models and sensitivity analysis*

In order to investigate the relative importance of the prior assumptions (6)–(11) in determining final student rankings, several alternative models were examined by varying the prior assumptions and assessing consequent differences in posteriors.

One-component achievement index or two? Perhaps the most controversial assumption made in the baseline model is that student performance can be adequately summarized through a single achievement index.

In regard to this issue, I note that standard GPA measures and GPA-based class rank are themselves univariate quantities, and so replacing a flawed univariate summary of student performance with a less-flawed univariate summary measure seems entirely reasonable. Of course, the value of measuring student performance with any univariate quantity is often questioned, and it is worth examining the loss of information incurred through such summaries.

As a first step toward examining this question, the baseline model was expanded so that the prior distributions on student achievement indices were modeled as bivariate normal random variables having the form

$$\begin{pmatrix} X_i^1 \\ X_i^2 \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}\right). \tag{14}$$

The value of the prior correlation between components of the achievement indices was assumed to be $\rho = 0.43$, based on observed correlations between SAT math and verbal scores. Results obtained using a prior correlation of 0 are quite similar to those reported below for $\rho = 0.43$.

To incorporate the bivariate achievement index into the model for grade generation, each academic department listing in the Duke Registrars coding of courses was assigned a weight w_d, $0 \le w_d \le 1$. In the expanded model, student i is assumed to receive a grade of k in class j in department listing d whenever

$$\gamma_{k-1}^j < w_{d(j)} X_i^1 + (1 - w_{d(j)}) X_i^2 \le \gamma_k^j, \tag{15}$$

where $w_{d(j)}$ denotes the department offering course j. Beta priors proportional to $w_d(1 - w_d)$ were assumed for each department weight, except in the case of the undergraduate writing course requirement. In order to make the weights and the components of the student achievement indices identifiable, a beta density proportional to $w_d^9(1 - w_d)$ was assigned to the department weight for the undergraduate writing course, which is taken by nearly all incoming Duke students.

Given the department weighting $w_{d(j)}$ for a particular class, the marginal distribution of student

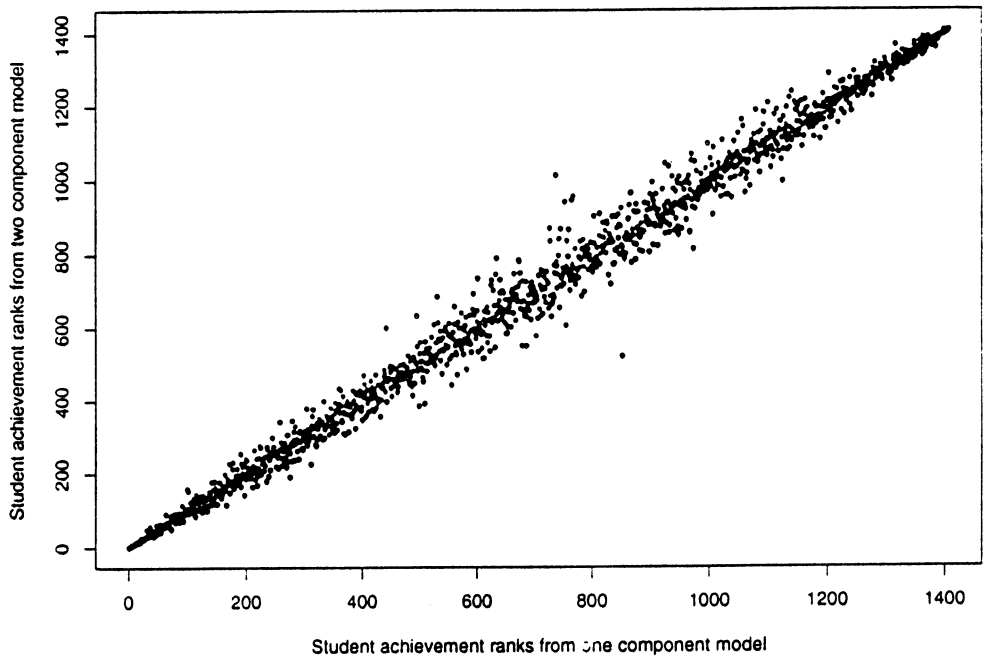

FIG. 6. *Rank of student achievement indices under one- and two-component models.*

achievement indices within a class is no longer $N(0, 1)$, but is instead $N(0, \sigma^2_{d(j)})$, where

$$\sigma^2_{d(j)} = w^2_{d(j)} + 2\rho w_{d(j)}(1 - w_{d(j)}) + (1 - w_{d(j)})^2.$$

To reflect this change, the prior distribution on the grade-cutoffs in the two-index model, given ι and again motivated by the Jeffrey's prior for assignment to observed categories, becomes

$$\begin{aligned} p(\boldsymbol{\gamma}^j \mid \iota) \propto{}& \Phi(\gamma_{I_{\min}}/\sigma_{d(j)})^{-1/2} \\ &\cdot \left\{ \prod_{\substack{i_k \in I \\ i_k > I_{\min}}} \left(\Phi\left(\frac{\gamma^j_{i_k}}{\sigma_{d(j)}}\right) - \Phi\left(\frac{\gamma^j_{i_{k-1}}}{\sigma_{d(j)}}\right)\right)\right\}^{-1/2} \\ &\cdot \left\{ \prod_{\substack{i_k \in I \\ i_k < I_{\max}}} \frac{\phi(\gamma^j_{i_k}/\sigma_{d(j)})}{\sigma_{d(j)}} \right\}. \end{aligned} \tag{16}$$

The bivariate student achievement indices obtained from a suitably modified ICM procedure were compared to those obtained in the baseline model using the criteria described above. Class ranks were computed for each student from the bivariate index by weighting the two components assigned to each student according to that student's selection of courses. The weighted achievement index for student i was thus defined as

$$X^w_i = \frac{1}{|\mathscr{C}_i|} \sum_{j \in \mathscr{C}_i} \left(w_{d(j)} X^1_i + (1 - w_{d(j)}) X^2_i\right). \tag{17}$$

Class ranks based on $\{X^w_i\}$ are plotted in Figure 6 against the corresponding ranks obtained from the baseline model. As illustrated, the correlation between the two rankings is quite high (0.996), and for all but two or three students the difference between class rank computed under the two models is relatively small.

The weighted achievement indices were also regressed on high school GPA and verbal and math SAT scores. The multiple R^2 for this regression was 0.352, somewhat higher than that obtained for the baseline model. In addition, the proportion of orderings correctly predicted for students enrolled in common classes under the two-component model was 0.148 (using estimated weights and achievement components appropriate for each class), which was just 0.017 units above the ideal rate of 0.131. Recall that for the baseline index the error rate was 0.168, or 0.037 units above the best obtainable rate.

In terms of accuracy of ranking, the two-component achievement index clearly outperformed the baseline index. However, the gains realized with the two-component index are offset by the additional complexity involved in explaining the two-component index to students, employers, college administrators and faculty. For this reason, the proposal now before the Arts and Sciences Council at Duke University (see Section 6) is based on the one-component index, though multicomponent models continue to be investigated. Further understanding of multicomponent models, in conjunction with an increased "comfort level" with

the baseline model, may make it possible to introduce multicomponent models for adoption at Duke University in the future. Of course, evaluating the number of components needed to describe student performance adequately, and correlating these components with specific academic fields, is itself an interesting topic and one on which I hope to report in a future article.

Normality of student achievement indices. Another important assumption made within the baseline model is that student achievement indices are distributed a priori according to a standard normal distribution. In reality, a better prior model for student achievement indices, given that intraclass variation is assumed to follow a normal distribution, would likely involve a mixture distribution on the marginal distribution on achievement indices. However, if such a model were employed, achievement indices would be shrunk toward different values on the achievement scale, depending on the mixture component to which they fell closest. By positing a standard normal distribution on the achievement indices, the prior distribution effectively shrinks all indices toward a common value of 0, which seems more palatable than shrinking student achievement indices towards disparate values. Furthermore, should a more complicated prior model be assumed for the achievement indices, it would likely not have a significant effect on the ranking of students, only the spacing between their estimated indices.

Empirical Bayes estimation of prior densities. The remaining model assumptions concern the particular inverse gamma distribution employed as the prior density on the instructor variances and the form of the prior density specified for the grade-cutoffs. To investigate the impact of these assumptions on final student ranking, alternative estimates of achievement indices were obtained using empirical Bayes methodology and more standard "uniform" priors.

From an empirical Bayes viewpoint, a prior density on the distribution of grade-cutoffs can be obtained by letting $\boldsymbol{\pi}_j$ denote the probability vector describing the multinomial probabilities that students in class j are assigned to different grade categories, and assuming that $\boldsymbol{\pi}_j$ is drawn from a Dirichlet distribution with parameters $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_K)$. That is, assume that

$$p(\boldsymbol{\pi}_j) \propto \prod_{1}^{K} (\pi_k^j)^{\omega_k - 1}, \tag{18}$$

independently in j. Equation (18) provides an achievement-index-free model for each class's multinomial probability vector describing grade assignments. Using the observed number of grades assigned in all classes, it is straightforward to obtain the maximum likelihood estimate for $\boldsymbol{\omega}$, denoted here by $\hat{\boldsymbol{\omega}}$. Transforming to the standard normal scale, the resulting empirical Bayes prior on grade-cutoffs is

$$p(\boldsymbol{\gamma}^j) \propto \left\{ \prod_{i=1}^{K} (\Phi(\gamma_{i_k}^j) - \Phi(\gamma_{i_{k-1}}^j)) \right\}^{\hat{\omega}_k - 1} \cdot \left\{ \prod_{i=1}^{K-1} \phi(\gamma_{i_k}^j) \right\}. \tag{19}$$

With this prior on grade-cutoffs and the standard normal prior on the achievement indices, it is possible to generate sample transcripts of students for any chosen value of instructor variance parameters (α, λ). By simulating student transcripts, an empirical Bayes method-of-moments approach can also be used for estimating values of the instructor variance parameters. More specifically, both the sample mean of the within-student variance of assigned grades and the variance of this mean student variance between students can be calculated directly from sampled student transcripts. By appropriate choice of α and λ, sample transcripts can be generated to mimic the observed values of these quantities. Implementing this moment-matching procedure yielded approximate values for α and λ of 6.0 and 5.7, respectively.

The sample correlation between the ICM estimates of student ranks based on achievement indices estimated from the model employing empirical Bayes priors and the baseline model was quite high—0.986. However, for a small minority of students the differences in class rank were as great as 20%, which again raises the question of which model should be preferred. For the empirical Bayes estimates, the multiple R^2 of the regression of the indices on SAT and high school GPA was 0.302, and the error in predicting the grades of paired students in common classes was 0.170. Both measures suggest that the baseline model was the more accurate, and further evidence of this assertion was obtained through residual analyses of the type performed for the baseline model.

The comparatively poor performance of the empirical Bayes estimates of student achievement is somewhat surprising, although a partial explanation for this failure may be found by examining the relationship between grading policies of instructors and classroom attributes. For example, a scatterplot depicting the relationship between class size and mean grade is provided in Figure 7, and suggests that the failure of the empirical Bayes model

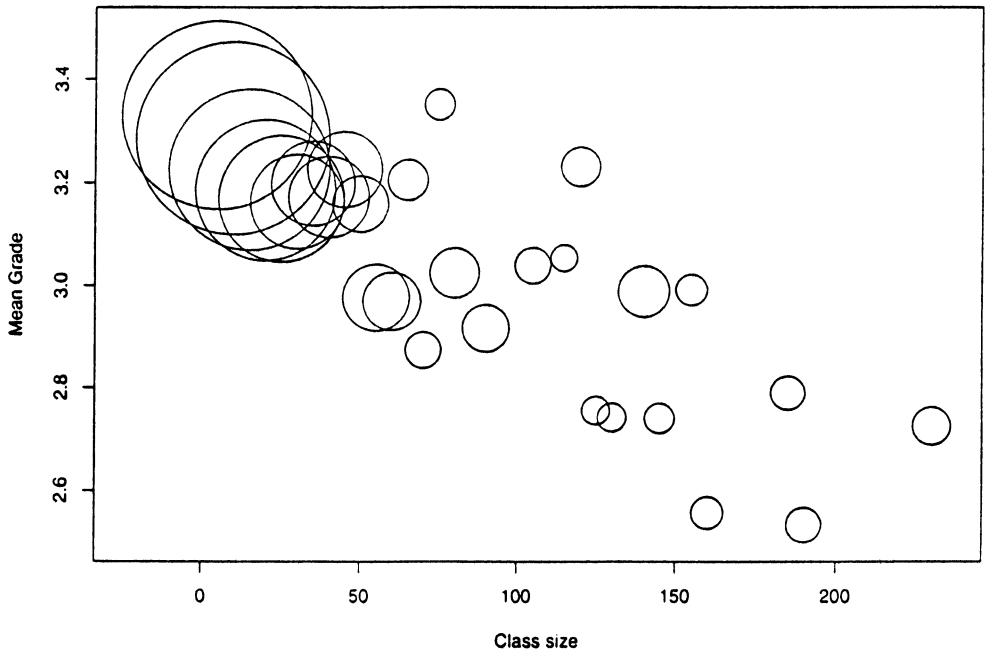

FIG. 7. *Plot of mean grade assigned versus number of students in class; circles in the plot have radii proportional to the square root of the number of students used to compute the mean grade. Class sizes were grouped in units of five students. That is, the leftmost circle represents the mean classroom grade of classes containing between* 1 *and* 5 *students, the next circle classes of size* 6 *to* 10 *and so on.*

might be caused in part by its failure to account for decreasing trends in mean grade with class size. It is likely that similar trends exist with other course attributes. Apparently, partially misspecified empirical Bayes priors are less effective in modeling grade generation than the relatively vague prior employed in the baseline model. These observations support the hypothesis that interval grade information cannot easily be incorporated into student assessments.

Locally uniform priors. Locally uniform priors can be stipulated for both the grade-cutoffs and the instructor-variance hyperparameters α and λ. Unfortunately, independent uniform priors on α and λ cause the posterior distribution to take its maximum value near the origin, making ICM estimation futile.

As an alternative to full ICM estimation for such models, one may instead fix the values of the variance hyperparameters near their posterior mean, and maximize the remaining model parameters as before. The posterior mean of (α, λ) can be approximated in a straightforward way by modifying the BDA/MCMC algorithm so that values of these parameters are sampled along with all others.

Two models employing locally uniform priors were studied. In the first, denoted U1, the grade-cutoffs were assumed to have the prior distributions specified in (11), but independent uniform priors were assumed for both α and λ. In the second model, U2, uniform priors on both the grade-cutoffs (subject to the ordering constraint $\gamma_k^j \leq \gamma_{k+1}^j$), and the hyperparameters α and λ were employed. For numerical stability, grade-cutoffs were assumed to be uniformly distributed within the interval $(-10, 10)$.

The posterior mean of (α, λ) under these two models were $(3.8, 4.6)$ (U1) and $(9.7, 2.8)$ (U2).

Interestingly, the inverse gamma prior employed in model U2 led to a posterior that concentrated its mass closer to 0 than did the other models considered. This was due to the interaction between the priors on the achievement indices and the grade-cutoffs. Because of the nonnegligible posterior probability assigned to the unobserved grade categories under model U2, A and B categories were forced to take cutoffs well above 0, while the $N(0, 1)$ prior on the achievement indices pulled the achievement indices toward 0. As a consequence, the posterior mean of achievement indices was approximately 0.8, and the posterior variance were 0.2. Instructor variance parameters decreased accordingly. In contrast, the posterior mean and variance of the achievement indices under model U1 were 0.2 and 1.2, respectively.

The multiple R^2 statistics for U1 and U2 were 0.342 and 0.351; the prediction errors for paired stu-

dents and classes were 0.168 and 0.171. The correlations of the student ranks obtained from models U1 and U2 with the baseline model's ranks exceeded 0.999 and 0.998, respectively.

The high correlation of ranks obtained under the various models of this section indicate that estimates of student achievement are robust to minor variations in model assumptions. Indeed, differences in student ranks obtained using these differing model assumptions is small compared to the posterior uncertainty of the same ranks obtained using any particular model. This point is illustrated in Figure 8, which depicts the posterior standard deviation of each student rank from the Duke cohort in the baseline model versus the absolute difference of ranks that were obtained using the baseline model and model U2. This figure demonstrates that the posterior standard deviation associated with each student's rank is larger, and in most cases much larger, than the difference attributable to model specification. In fact, on average the two ranks obtained from the baseline model and model U2 were less than 1/5 of a standard deviation apart. This shows that the posterior variance associated with achievement indices dwarfs the uncertainty associated with the specific choice of prior hyperparameters and that the relatively vague priors chosen in the baseline model likely provide nearly optimal estimates of class rank among one-component models.

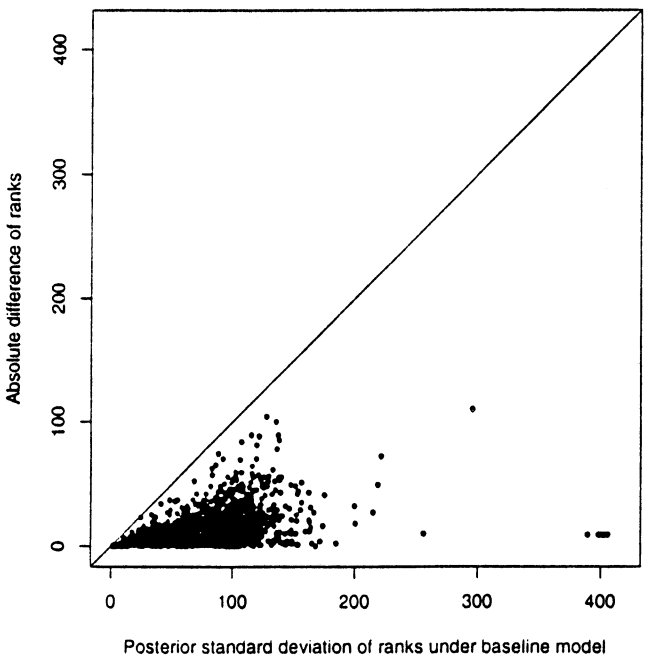

FIG. 8. *Posterior variability of AI-based class rank: this figure depicts the posterior standard deviation of the rank of each Duke student under the baseline model versus the absolute difference of the ICM ranks obtained under the baseline model and model U2. The area under the line contains those points at which the differences in rank were smaller than the posterior standard deviation of the rank under the baseline model. No points fell above the line.*

5. DISCUSSION

From a technical standpoint, the primary innovations of the Bayesian model over previously proposed GRM-type models are the introduction of prior distributions on instructor variances and grade-cutoffs, and the use of the binary random variables $\boldsymbol{\iota}$ to model unobserved grade categories. The priors on the instructor variance parameters and grade-cutoffs replace the hard constraints on grade-cutoffs employed in earlier models and allow model parameters to adjust more accurately to observed grading patterns. Combined with the effect of the auxiliary variables $\boldsymbol{\iota}$, the proposed model for grades is essentially able to discard all extraneous interval information contained in observed grade data, and instead relies on only the rankings of students within classes. As a result, shifting all grades assigned to a class of students up or down has no effect on estimated student achievement indices. An instructor who assigns A's to every student in a class is no longer doing the students in his or her class a favor; such a grading policy has no effect on the achievement indices of any of the students in the class. Likewise, instructors who assign grades that are significantly below the university average no longer penalize students in their classes.

The proposed model offers educators an alternative to GPA-based methods for evaluating student performance. The critical features of this model are that it reduces the subjectivity associated with the interpretation of instructor grade assignments and largely eliminates incentives for students to enroll in less rigorous courses. Expected long-term effects of adopting this evaluation scheme include increased enrollments in upper-level undergraduate classes, increased enrollments in mathematics and natural science classes, reduced pressure on instructors to inflate grades and a greater desire on the part of faculty to reward excellence in the classroom through differential assignment of student grades.

6. POSTSCRIPT

A grade-reform proposal based on the method described in this article is currently being considered as a replacement for GPA for all undergraduate students at Duke University. At printing, this proposal

TABLE 3

Course	Grade	Mean grade	Estimated achievement	Mean achievement	Mean GPA	Grade cutoffs
Student 1: GPA-based rank, 296; achievement index Rank, 822						
CHM 01––	B–	2.90	−0.42	0.02	3.23	(−0.61, −0.25)
MUS 05––	B+	3.74	−1.46	0.00	3.46	(−9.99, −0.89)
PHL 04––	B	2.88	−0.58	−0.36	3.26	(−0.86, −0.31)
UWC 00––	B+	3.72	−2.01	0.04	3.30	(−9.99, −1.25)
BIO 04––	A	4.00	−0.23	0.33	3.43	(−9.99, 9.99)
CHM 01––	B	2.95	0.00	0.08	3.25	(−0.26, 0.30)
ECO 05––	B+	2.86	0.36	0.16	3.35	(0.16, 0.57)
MTH 03––	B–	2.67	−0.27	−0.02	3.23	(−0.35, −0.20)
ECO 14––	B+	3.30	−0.23	−0.23	3.59	(−9.99, 9.99)
GER 01––	A	3.85	1.11	−0.13	3.29	(−0.16, 9.99)
STA 11––	A–	3.28	0.68	0.05	3.33	(0.43, 1.00)
CPS 01––	A	3.71	−0.05	−0.29	3.22	(−0.59, 0.56)
CST 14––	A	3.04	1.65	−0.18	3.24	(0.95, 9.99)
ECO 14––	B	3.47	−1.21	0.22	3.45	(−9.99, −0.35)
GER 06––	A–	3.80	−0.36	0.51	3.49	(−9.99, 0.94)
PE 10––	A	4.00	−0.62	−0.34	3.14	(−9.99, 1.05)
ECO 06––	A–	3.70	−0.23	−0.23	3.59	(−9.99, 9.99)
GER 11––	A	4.00	−0.23	−0.23	3.59	(−9.99, 9.99)
GER 11––	A–	3.70	−0.23	−0.23	3.59	(−9.99, 9.99)
PS 10––	A–	3.70	−0.23	−0.23	3.59	(−9.99, 9.99)
GER 15––	A	4.00	−0.23	−0.23	3.59	(−9.99, 9.99)
GER 15––	A	4.00	−0.23	−0.23	3.59	(−9.99, 9.99)
HST 10––	A	4.00	−0.23	−0.23	3.59	(−9.99, 9.99)
PS 10––	A	4.00	−0.23	−0.23	3.59	(−9.99, 9.99)
	3.59	3.55	−0.23	−0.08	3.41	
Student 2: GPA-based rank, 822; achievement index rank, 350						
CHM 01––	A+	2.90	2.54	0.02	3.23	(2.38, 9.99)
LAT 06––	A	3.65	2.01	0.65	3.46	(0.42, 9.99)
MTH 10––	A	2.76	2.01	0.79	3.48	(1.59, 9.99)
UWC 00––	A	3.27	1.66	−0.69	2.97	(0.97, 9.99)
BIO 12––	B+	3.10	0.48	0.63	3.44	(−0.15, 1.08)
CHM 01––	A	2.95	1.44	0.08	3.25	(1.11, 2.56)
LAT 06––	B+	3.57	−1.36	0.17	3.22	(−9.99, −0.47)
MTH 10––	C+	3.14	0.04	1.05	3.61	(−0.34, 0.32)
CHM 15––	A	2.83	0.86	0.09	3.12	(0.09, 9.99)
CA 14––	A–	3.70	0.62	0.62	3.21	(−9.99, 9.99)
LAT 10––	B	3.00	0.16	0.07	3.12	(−0.37, 0.61)
PHY 05––	B–	2.85	0.17	0.37	3.35	(0.05, 0.28)
STA 11––	D	2.47	−1.16	0.24	3.25	(−9.99, −0.74)
BIO 16––	D+	2.98	−1.67	0.48	3.36	(−9.99, −1.44)
CA 14––	C+	2.30	0.62	0.62	3.21	(−9.99, 9.99)
LAT 10––	D–	1.00	0.62	0.62	3.21	(−9.99, 9.99)
PHY 05––	D	2.56	−1.09	0.51	3.39	(−9.99, −0.85)
ARB 00––	A	4.00	0.62	0.56	3.42	(−9.99, 9.99)
BIO 15––	B–	2.86	0.28	0.41	3.32	(0.03, 0.51)
BIO 18––	A	2.62	1.85	0.36	3.24	(1.47, 9.99)
SP 00––	B	3.00	0.62	0.62	3.21	(−9.99, 9.99)
ARB 00––	A	4.00	0.62	0.62	3.21	(−9.99, 9.99)
BIO 19––	A	4.00	0.62	0.62	3.21	(−9.99, 9.99)
REL 15––	A+	3.74	1.65	−0.09	3.25	(1.00, 9.99)
SP 00––	A–	3.37	1.14	.35	3.36	(−0.32, 9.99)
BIO 15––	A	3.89	1.45	.44	3.38	(−0.04, 9.99)
BIO 27––	A–	3.70	0.62	.62	3.21	(−9.99, 9.99)
ARB 06––	A–	3.70	0.62	.62	3.21	(−9.99, 9.99)
ARB 19––	A	4.00	0.62	.62	3.21	(−9.99, 9.99)
BIO 22––	A	3.43	1.12	.46	3.41	(0.67, 9.99)
SP 06––	B+	3.80	−1.44	.10	3.33	(−9.99, −0.79)
ARB 19––	B	3.50	−0.57	.64	3.35	(−9.99, 0.50)
BIO 19––	C–	1.70	0.62	.62	3.21	(−9.99, 9.99)
SP 07––	C–	2.20	−0.50	.18	3.07	(−9.99, 0.15)
SP 13––	A	4.00	0.62	.62	3.21	(−9.99, 9.99)
SP 13––	A	4.00	0.62	.62	3.21	(−9.99, 9.99)
	3.21	3.18	0.62	0.42	3.27	

had gained unanimous approval of the Committee on Grades, a subcommittee of the Academic Affairs Committee, itself a committee of the Arts and Sciences Council. With the exception of a student representative, the Academic Affairs Committee also unanimously approved the proposal. A vote before the Arts and Sciences Council is scheduled for early 1997. Should the proposal pass the Arts and Sciences Council, a five-year phase-in of the achievement index would begin in the 1997–1998 academic year.

In the first year of phase-in, the proposal stipulates that undergraduates at Duke will receive a letter at the end of each semester informing them of their AI-based class rank and an AI-adjusted GPA. The adjusted GPA will be obtained by matching each student's achievement index with the corresponding percentile of the observed, unadjusted GPA from that class. By scaling the AI-adjusted GPA in this way, the distribution of AI-adjusted GPA is identical to the distribution of unadjusted GPA. This fact appears to be very important to students, who are concerned that use of the AI-adjusted GPA will affect their chances of entering professional schools and graduate schools. Under this proposal, the interpretation of AI-adjusted GPA for admissions officers should be largely transparent. A short explanation of the AI-rank and AI-GPA will accompany each letter, and a more detailed explanation will be offered electronically (a draft web site containing this information currently resides at http://www.phy.duke.edu/~gauthier/grades/grades.html). In the remaining years of the phase-in, AI-based class rank and AI-GPA will be reported on student transcripts, along with unadjusted GPA and GPA-based class rank. A brief summary of the achievement index will appear on student transcripts, and Duke University will make a concerted effort to educate graduate and professional schools on the interpretation of the achievement index values. Following yearly reviews of the reform during the phase-in, GPA and GPA-based class rank will be removed from student transcripts in the sixth year of implementation.

The expressed goals of the Academic Affairs Committee in recommending this reform are to eliminate the inequities inherent in unadjusted GPA measures; to reduce the increasingly common practice among students of selecting courses on the basis of expected grade; to combat grade inflation by encouraging faculty to assign grades differentially on the basis of student performance; and to further enhance the intellectual environment within the Duke University community.

APPENDIX: SELECTED TRANSCRIPTS OF DUKE UNIVERSITY UNDERGRADUATES

The columns in the transcripts shown in Table 3 represent (1) course designator, (2) grade, (3) mean grade in class, (4) estimated classroom achievement index of the student, (5) mean achievement index of all students enrolled in the class, (6) mean GPA of students in the class and (7) estimated category cutoffs for the grade received. Approximately 1,400 students were ranked.

ACKNOWLEDGMENTS

The author thanks Richard White, Dean of Trinity College and Vice Provost for Undergraduate Education, and Harry DeMik, Deputy University Registrar, for assistance in obtaining the undergraduate records studied in this article. The author also thanks Professors Daniel Graham and Daniel Gauthier, members of the Committee on Grades, for their support of a proposal based on this methodology as a replacement for GPA at Duke University. Finally, I would like to acknowledge the contribution of four anonymous referees whose comments significantly improved this manuscript.

REFERENCES

ALBERT, J. H. and CHIB, S. (1993). Bayesian analysis of binary and polychotomous response data. *J. Amer. Statist. Assoc.* **88** 669–679.

ALBERT, J. H. and CHIB, S. (1995). Bayesian residual analysis for binary response regression models. *Biometrika* **82** 747–759.

BESAG, J. E. (1986). On the statistical analysis of dirty pictures. *J. Roy. Statist. Soc. Ser. B* **48** 259–302.

BRADLOW, E. T. (1994). Analysis of ordinal survey data with "no answer" responses. Ph.D. dissertation, Dept. Statistics, Harvard Univ.

BRADLOW, E. T. and ZASLAVSKY, A. M. (1997). A hierarchical latent variable model for ordinal data with "no answer" responses. Preprint.

CAULKIN, J., LARKEY, P. and WEI, J. (1996). Adjusting GPA to reflect course difficulty. Working paper, Heinz School of Public Policy and Management, Carnegie Mellon Univ.

COWLES, M. K., CARLIN, B. P. and CONNETT, J. E. (1996). Bayesian Tobit modeling of longitudinal ordinal clinical trial compliance data with nonignorable missingness. *J. Amer. Statist. Assoc.* **91** 86–98.

ELLIOTT, R. and STRENTA, A. (1988). Effects of improving the reliability of the GPA on prediction generally and on comparative predictions for gender and race particularly. *Journal of Educational Measurement* **25** 333–347.

FISCHER, G. H. and MOLENAAR, I. W., eds. (1995). *Rasch Models: Foundations, Recent Developments, and Applications*. Springer, New York.

GELFAND, A. E. (1996). Model determination using sampling-based methods. In *Markov Chain Monte Carlo in Practice* (W. Gilks, S. Richardson and D. J. Spiegelhalter, eds.) 145–161. Chapman and Hall, London.

GILKS, W., RICHARDSON, S. and SPIEGELHALTER, D. J. (1996). Introducing Markov chain Monte Carlo. In *Markov Chain*

Monte Carlo in Practice (W. Gilks, S. Richardson and D. J. Spiegelhalter, eds.) 1–20. Chapman and Hall, London.

GOLDMAN, R., SCHMIDT, D., HEWITT, B. and FISHER, R. (1974). Grading practices in different major fields. *American Education Research Journal* **11** 343–357.

GOLDMAN, R. and WIDAWSKI, M. (1976). A within-subjects technique for comparing college grading standards: implications in the validity of the evaluation of college achievement. *Educational and Psychological Measurement* **36** 381–390.

JOHNSON, V. E. (1996). On Bayesian analysis of multirater ordinal data. *J. Amer. Statist. Assoc.* **91** 42–51.

LARKEY, P. (Jan. 25, 1991). A better way to find the top scorer. *Golf World* 72–74.

LARKEY, P. and CAULKIN, J. (1992). Incentives to fail. Working Paper 92-51, Heinz School of Public Policy and Management, Carnegie Mellon Univ.

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

LORD, F. M. and NOVICK, M. R. (1968). *Statistical Theories of Mental Test Scores*. Addison-Wesley, Reading, MA.

MURAKI, E. (1990). Fitting a polytomous item response model to Likert-type data. *Applied Psychological Measurement* **14** 59–71.

NANDRAM, B. and CHEN, M.-H. (1996). Accelerating Gibbs sampler convergence in the generalized linear models via a reparameterization. *J. Statist. Comput. Simulation* **45**. To appear.

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

STRENTA, A. and ELLIOTT, R. (1987). Differential grading standards revisited. *Journal of Educational Measurement* **24** 281–291.

TANNER, M. and WONG, W. H. (1987). The calculation of posterior distributions by data augmentation. *J. Amer. Statist. Assoc.* **82** 528–549.

VAN DER LINDEN, W. J. and HAMBLETON, R. K., eds. (1997). *Handbook of Modern Item Response Theory*. Springer, New York.

WANG, X., WAINER, H. and THISSEN, D. (1995). On the viability of some untestable assumptions in equating exams that allow examinee choice. *Applied Measurement in Education* **8** 211–225.

YOUNG, J. W. (1989). Developing a universal scale for grades: investigating predictive validity in college admissions. Ph.D. dissertation, School of Education, Stanford Univ.

YOUNG, J. W. (1990). Adjusting the cumulative GPA using item response theory. *Journal of Educational Measurement* **12** 175–186.

YOUNG, J. W. (1992). A general linear model approach to adjusting cumulative GPA. *Journal of Research in Education* **2** 31–37.

YOUNG, J. W. (1993). Grade adjustment methods. *Review of Educational Research* **63** 151–165.

Comment: Adjusting Grades at Duke University

Patrick D. Larkey

The Duke University project described in Valen Johnson's paper, "An alternative to traditional GPA for evaluating student performance," was very exciting because it was the first serious attempt to diagnose and solve important problems with extant grading practices. Unfortunately, since Johnson's paper was completed, the Arts and Sciences Council at Duke voted 19 to 14 against experimenting with the "achievement index" in parallel with traditional GPA. The opposition apparently came from the social sciences and humanities professors on the Council (Gose, 1997). The initiative is apparently dead at Duke for the moment.

Patrick D. Larkey is Professor of Public Policy and Decision Making, H. J. Heinz III School of Public Policy and Management, Carnegie Mellon University, Hamburg Hall 242, Pittsburgh, Pennsylvania 15213 (e-mail: pl15@andrew.cmu.edu).

The issue will not, however, go away in higher education because the incentives problems with current grading practices are very serious, because the technical means for solving the problems exist and because it will prove impossible to maintain, much less improve, the quality of education with the prevailing incentives. Grading practices in conjunction with other flawed performance measures (e.g., appraising the quality of courses and instructors primarily, if not exclusively, through the responses of a subset of students who have just sat through 80% of a course) are responsible for a lot of mischief in higher education.

There are, however, significant obstacles to replacing GPA. The first important obstacle is that the full extent of the problem has not been established. Obviously GPA is wrong (see Johnson, Figure 1); some A's are harder for most students to get than other A's. Obviously students, to the extent that they are motivated by GPA rather than substantive consider-

ations, have incentives to avoid courses and instructors where greater skill and/or effort is required to achieve a satisfactory grade or where the grading is simply below the institutional norm. However, we do not yet have the careful empirical studies of how students choose courses to know the full extent of the problem. The casual evidence is compelling. Many students freely admit that the difficulty of courses and instructors is an important consideration in selecting their schedule. At Duke, for example, one student noted that adoption of the achievement index might have induced him to take quantum mechanics, one of the more difficult courses (Gose, 1997).

The second obstacle is communicating the diagnosis and the designed solution to audiences including members who have difficulty understanding analytic arguments and whose interests may be to not understand them. The source of opposition at Duke is not surprising. The few studies that have been done all indicate that there has been relatively more grade inflation in "softer" subjects. While there has been inflation in the physical sciences and mathematics, there is apparently more resistance to inflation in domains with more sharply and logically defined right and wrong answers. Grade inflation has been an important edge for some fields in competing for students as core curricula have waned and student choices have waxed. It is a perverse form of price competition; they have been able to offer higher grades for equivalent or lesser amounts of work. Not incidentally, they also get higher levels of student satisfaction and better external appraisals of their "teaching quality" which increasingly influence promotions and salary adjustments in elite research institutions.

The general analytic problem of which correcting grade point average is a specific instance is:

> Rank competitors in terms of their performance across nonequivalent tournaments where individual competitors compete in different, nonrandom subsets of all possible tournaments.

This problem is surprisingly common. It is found in professional golf (Larkey, 1991), airline performance evaluations (Caulkins et al., 1993) and a variety of other situations where there is a need to compare dissimilar performers performing in dissimilar circumstances. Any large organization, for example, hiring new employees with a procedure that does not allow all interviewers to see all interviewees or assign interviewees to interviewers randomly should not aggregate and/or average interview scores without some correction for differences among interviewers.

The general consequences of not recognizing and solving this problem are (1) the rankings are improper and whatever rewards and punishments may attach to rank are improperly assigned and (2) where competitors have discretion in choosing the tournaments in which they will compete, they have incentives to choose strategically rather than improving their performance on the proverbial "level playing field" to improve their rank. In professional golf the top 125 players on the uncorrected money list at the end of each year get a full playing exemption—the right to play in any of the regular tour events they choose—in the next year. In the airline industry, some airlines (e.g., Alaska Airlines) that are superior to their competitors but that are disadvantaged by the subset of airports they frequent are shown to be inferior in terms of on-time statistics because of the uncorrected measures of on-time performance widely reported by the Federal Aviation Administration.

Nowhere are the specific consequences of not recognizing and solving this problem more serious than in grading. Professional golfers and airlines are probably not strongly motivated to choose their subsets of "tournaments" to maximize the improper performance measure. Students probably are.

The improper incentives for students in choosing programs and courses are the most important consequence of the flawed grading practices rather than grade inflation. Grade inflation results from a larger dynamic in which the flawed grading practices interact with flawed course or instructor evaluations, namely, a positive correlation between more generous grading and levels of student satisfaction.

The primary obstacle to solving the grade point average problem is not technical. There are many possible measures of aggregate performance for comparing students that are superior to GPA. They are superior in that they better represent comparative performance and remove incentives for students to choose courses and instructors based on their relative difficulty.

Given multiple alternatives to GPA and the absence of any source for "correct" measures of student performance, model choice is a serious problem. There are two bases for preferring one model to another. First, we can compare the face validity of the resulting rankings. This may be easiest to do on contrived problems, complex versions of Johnson's Figure 1. Second, we can appraise the logic of the models. This approach is probably not conclusive when comparing models that both provide max-

imum likelihood estimates of how students would have performed in courses they did not take. Given the difficulty of the intended audience, the ease of communicating the model should also play an important role; simplicity is probably much more important than technical virtuosity if the models provide roughly equivalent results.

Faculty members are often ill-equipped to understand the problem. Many faculty members had very little difficulty with courses over their student careers. Many had sufficient capacity relative to their student peers to choose courses on the basis of career interests and personal development rather than on the basis of their relative difficulty. Faculty find it difficult to understand, much less empathize with, the less capable or lazy students who actively strategize on how to survive the curriculum rather than how to benefit from it.

One department head at Duke who opposed the GPA correction initiative is reported to have said, "Unless you understand what is driving the grades up, you shouldn't be driving the grades down." This is a fundamental misunderstanding of what was proposed. The proposal was to reinterpret and redistribute grades. Some would be higher and others lower than the unit weight in the traditional GPA assignment.

A key element in an eventually successful strategy to reform grading practices must be to get the primary consumers of grade information to insist on better information about student performance than grades provide. Professional and graduate schools, all manner of employers, and parents are the primary consumers. Many of the schools and employers have already adapted to the increasing meaninglessness of grades over the past 20 to 30 years. There is greater reliance on the overall reputational quality of the institution (the selection effect), on performance in specific courses where grades can be calibrated and on personal interviews. Each of these adaptations is flawed. The collection is no substitute for meaningful information from the most complete sample on performance, all courses in an undergraduate or graduate career.

Duke has raised consciousness on an important set of issues. Now we need the thorough empirical work and clever marketing strategies that will lead to the needed reforms.

Comment: Grade Inflation, A Pervasive Problem—A Commentary on Johnson's Achievement Index

John W. Young

GRADE INFLATION

Grade inflation is, at present, a pervasive problem in secondary schools and in higher-education institutions throughout the country. Data in the form of anecdotes, local surveys and national studies have documented the relentless upward trend of grades during the past three decades. None of this would be of concern if other indicators of student achievement or performance also demonstrated consistent increases over time. Unfortunately, other external indicators such as national standardized test scores do not show a corresponding rise when grade inflation has been most evident. Historically, the study of grade inflation as an educational phenomenon and attempts to eliminate it date back as far as the beginning of the 20th century (Young, 1993).

What is the evidence on grade inflation? At the high school level, more than twice (32% versus 15%) as many first-year college students reported in a 1996 national survey that their overall high school grade point average (GPA) was an A-minus or better as compared with 30 years earlier (Hornblower, 1997). Yet if students are doing better in schools, why are there now more complaints about the poor quality of schools? In many school districts, it is not uncommon for top-ranking students to flaunt GPA's of 4.2 on a grade scale that supposedly ranges from 0 to 4. One has to wonder what interpretation

John W. Young is Associate Professor, Graduate School of Education, Rutgers University, 10 Seminary Place, New Brunswick, New Jersey 08903 (e-mail: jwyoung@rci.rutgers.edu).

to attach to such GPA's. GPA's over 4 are possible because grades in honors or advanced placement courses are calculated using a multiplier greater than 1.

The results from national tests do not mirror the rise of grades in schools. Scores from the National Assessment of Educational Progress, a biennial assessment of student achievement in elementary and secondary schools mandated by Congress, have been constant in reading proficiency from 1971 to 1992 and have shown only modest increases in mathematics proficiency (National Center for Education Statistics, 1995). Scholastic Assessment Test (SAT) scores declined over a 20-year period starting in the late 1960's and have risen only slightly since. Longitudinal changes in SAT scores are more difficult to interpret due to demographic changes in the composition of the test-taking population and because of self-selection factors in choosing to take the test. However, it is clear that the trend in SAT scores has not been commensurate with the corresponding rise in grades. Furthermore, because of less rigorous grading standards and grade inflation in high schools, the advantage of high school GPA over SAT scores in predicting college GPA has essentially vanished (Bejar and Blew, 1981). The correlation of high school GPA with college GPA has declined while the correlation of SAT scores with college GPA has held constant so that both are now equally good predictors.

Data from colleges and universities are no less convincing about grade inflation. For example, the average undergraduate grade at Harvard rose from the midpoint between a B− and a B to better than a B+ during the 25-year period from 1967 to 1992 without a corresponding rise in test scores (Lambert, 1993). In addition, grades in the humanities at Harvard rose much more than in the social or natural sciences. The so-called gentleman's C is now at least the "gentleman's B" if not higher. Other first-tier universities, such as Stanford, report equally dramatic increases in the average grade assigned. With continuing grade inflation, one unfortunate impact is that admissions to graduate and professional schools will be made primarily on test results rather than on grades since students can no longer be distinguished on the basis of grades.

Solutions to the problem of grade inflation such as the one developed by Professor Johnson are much needed. All of the pressures on grading are upward; remediating grade inflation cannot be accomplished individually but must be carried out collectively. Some institutions are attempting to deal with grade inflation by reinstating low grades; for example, after a long absence from students's transcripts, the F grade is again part of the Stanford grading system (Fetter, 1995). However, this approach is likely to have an impact only at the bottom end of the grade distribution without materially affecting the average grade assigned. An alternative approach at Dartmouth, begun two years ago, is to report the median course grade along with the student's grade so that some context is available for interpreting student transcripts. Though worthy of consideration, both the Stanford and Dartmouth solutions are crude instruments being applied globally.

The implementation of Johnson's achievement index at Duke will lead to fairer comparisons among students. Grading differences among instructors, courses and departments will impact students's GPA's to a much lesser degree than is true now. Students who major in departments with rigorous grading standards (typically, in the natural and mathematical sciences) will not be at so large a disadvantage in the competition for employment, for admissions to graduate school and for fellowships and scholarships. Furthermore, the incentive system for instructors will be converted to one of assigning high grades to deserving students in a course rather than the present system of assigning high grades to everyone.

Grade inflation is a serious and pervasive problem throughout education today. In the course of time, we will learn if Johnson's achievement index has served its purpose or whether further refinements are necessary, but this solution moves us in the right direction. In conclusion, Professor Johnson is to be commended for developing his achievement index and the faculty at Duke is to be praised for their courage in implementing it.

Comment: Achievement Index

Richard A. White

1. BACKGROUND: THE IMPETUS FOR DISCUSSION

During 1995–1996, Duke, along with other institutions across the country, engaged in dialogue concerning two issues central to the quality of undergraduate education: academic rigor in courses and grade inflation. This discussion arose from two initiatives which, it is important to note, arose separately and independently. The first was a concern expressed by the Provost, John Strohbehn, about grade inflation. Following campuswide discussion (and rejection) of the "Dartmouth model" of assigning a mean class grade beside letter grades on student transcripts, the Provost asked the College to consider how best to address this issue and respond to related issues of academic rigor. At the same time, Professor Valen Johnson of the Institute of Statistics and Decision Sciences developed and publicized a new concept for expressing student performance: the achievement index. Together, this peaked interest in the administration in examining grading practices at Duke and this innovative proposal for calibrating or recalibrating achievement made lively topics for a series of important campuswide discussions.

2. THE CONCEPT OF THE ACHIEVEMENT INDEX: THE PROCESS OF CONSIDERATION

During 1996–1997, Professor Johnson met with faculty and administrators and presented the principles of the AI to several key University committees, including the Academic Affairs Committee of the Arts and Sciences Council, the Undergraduate Administrative Group and the Directors of Undergraduate Studies within Trinity College. Augmenting these discussions were presentations and conversations with key student groups, including the Duke Student Government. As a result of these discussions, the primary focus of discussion shifted from concerns over grade inflation to concerns about "fairness" with regard to the GPA and rank in class; the traditional GPA evaluates students taking completely different kinds of courses as if all were equally rigorous and equally rigorously graded. Discussion about the introduction of the AI attempted to address several central questions: Do students choose courses on the basis of expected grading (so-called easy A courses)? Do students avoid risk-taking and fail to experiment with courses that are viewed to be extraordinarily rigorous or graded more rigorously or unknown? Is it fair to develop a GPA based upon such variation in courses (this is a mixture of rigorous courses, nonrigorous courses, technical or applied courses, experiential learning, cooperative grading)? Given the fact that class ranking is based upon GPA, is class rank under the current system a true indicator of academic excellence?

3. CURRENT STATUS AND DIRECTIONS: WHERE WE GO FROM HERE

Faculty in Arts and Sciences discussed the adoption of the AI at a "town meeting" in January 1997 to which all members of the University community were invited. In-depth discussion followed at the February meeting of the faculty council, with final vote to be taken on March 13. In these sessions, concerns have focused on such issues as how the AI-GPA would impact Duke students vis-à-vis other institutions? For example, will the AI-modified GPA or class rank handicap admission to professional schools and graduate school? While our information indicates that professional schools compute their own data directly from the transcript grades, we are making inquiries about how this proposal might impact admission to graduate schools and the use of the transcript by employers in the workforce.

In addition, two primary internal concerns have arisen. The first relates to independent study and small seminars, where many students may receive high grades (i.e., A's). As originally proposed, where there is no differentiation among the grades in these classes, grades received do not affect the AI. There is currently, however, a strong sense among the faculty and students that it is "unfair" for the high grades "not to count" and/or that there are no differences in the AI between all A-grade and all C-grade classes. Because better students may elect more independent studies and small seminars (and receive better

Richard A. White is Dean, Trinity College of Arts and Sciences, and Vice Provost for Undergraduate Education, Duke University, 104 Allen Building, Box 90042, Durham, North Carolina 27708-0042 (e-mail: rwhite@acpub.duke.edu).

grades), this lack of differentiation seems inappropriate. Because of this independent study/seminar issue, the algorithm will likely be modified.

Another issue that students have raised is a possible increase in competition among students within the same class. This notion seems to be driven by the perceived need for more rigorous grading, and some question how the altered classroom climate might work against individuals helping each other, working in small groups and team projects. An analysis, however, indicates that mentoring helps all students involved, and, therefore, the negative effect on the mentor may be largely offset.

4. CRITERIA FOR EVALUATION: MEASURES OF SUCCESS AND IMPLEMENTATION

If the Arts and Sciences Council adopts the AI "experiment," immediate attention must be paid to its evaluation and to the development of criteria to determine the AI's "success" or "failure." Such criteria, for example, might include patterns of course selection. Do students become more adventuresome in their course selection? Do they accept more "risk" and take challenging and diverse courses? Do students take more upper-level work?

According to current plans, we anticipate four years of transition between the current GPA system and a final decision of whether to adopt an AI-GPA alone. In the first year of the experiment, students will be informed individually of their AI-GPA by letter, but no formal indication will appear on their transcripts. Next, the AI may apply to incoming students, and current students will have the choice between the standard GPA and the AI-adjusted GPA. In subsequent years, both systems may appear, with a final determination to be decided at the end of the transition period.

Always in proposals of this sort, there are anticipated as well as unanticipated consequences. Although this proposal was not developed in direct response to concerns related to grade inflation, we expect that it may well have at least an indirect effect on this problem. We may see that faculty members may differentiate more rigorously among the students in the class in order for the grades to affect the AI, because, as noted above, where there is no differentiation, the grades do not impact the AI. There may be indirect pressure for both students and faculty to want a broader grade distribution within a class, and we look forward to the discussion and campus reaction with keen interest.

Note added in proof. On March 13, 1997, the Arts and Sciences Faculty Council voted not to pursue implementation of the Achievement Index at this time. Discussion will continue in the Fall Semester 1997.

Comment

Brian W. Junker and Eric T. Bradlow

Val has nicely synthesized two areas—modern applied Bayesian statistics and item response modeling—to provide a novel analysis in a third area that needs modern and sensible thinking: the evaluation and comparison of college students based on their reported performance in the courses they take. We have chosen to focus our discussion of his article on (1) how the model is motivated and (2) some technical choices and general comments. Our comments are intended to be friendly to Val's work, of which we think highly. We also take the opportunity to discuss in this forum (3) the richness of applications and interesting problems available to statisticians in educational statistics, which Val's research well exemplifies.

Brian W. Junker is Associate Professor, Department of Statistics, Carnegie Mellon University, 232 Baker Hall, Pittsburgh, Pennsylvania 15213 (e-mail: brian@stat.cmu.edu). Eric T. Bradlow is Assistant Professor of Marketing and Statistics, Wharton School of Business, University of Pennsylvania, Suite 1400 SH-DH, Philadelphia, Pennsylvania 19104-6371 (e-mail: ebradlow@wharton.upenn.edu).

1. MODEL MOTIVATION AND HISTORY

Val motivates his likelihood (5) for GPA adjustment by referring to the relative rankings of students within classes. An important fact omitted from Val's discussion is that, in addition to ranking

students on their estimated achievment X_i, the parameters γ_k^j in the model serve to locate the classes with respect to one another on the *same* scale as the X_i. Like other item response models (Fischer and Molenaar, 1995; van der Linden and Hambleton, 1997) this model simultaneously scales "subjects" (college students, in our case) and "stimuli" (the classes they take). This enables us to discuss not only the rankings of students within classes, but also the relative position of classes to each other and, most important, of students to classes—so that a high grade in an easy class may be more comparable to a low grade than a high one in a harder class. This is what makes Val's adjusted GPA more attractive from a fairness standpoint than the usual unadjusted GPA.

The model (5) has a long history in educational statistics and is currently in wide use. As Val notes, it is essentially Samejima's (1969, 1997) graded response model (GRM), modified to account for the fact that not all students take the same classes. Similar models can be found in statistical analyses of educational assessment surveys with complex survey designs, such as the National Assessment of Educational Progress (NAEP): see, for example, Algina (1992), Johnson, Mislevy and Thomas (1994), Patz (1996) and Zwick (1992). In fact, this identical model has appeared in research on general methods for ordinal data structures (e.g., grades) using latent variables in a Bayesian framework: see, for example, Albert and Chib (1993), Bradlow (1994) and Bradlow and Zaslavsky (1996). Hemker, Sijtsma, Molenaar and Junker (1997) provide a comparative review of such models that explores the relationship between total score (unadjusted GPA) rankings and latent variable (adjusted GPA) rankings in some generality.

2. SOME TECHNICAL ISSUES AND GENERAL COMMENTS

2.1 The Missingness Process

The fact that not all students take the same classes results in a great deal of missing data. Comparing students using GPA's (adjusted by Val's model or not) is an instance of what is called the *equating* problem in the educational statistics literature (see, e.g., Holland and Rubin, 1982). Equating problems also arise when SAT scores are compared from one year to the next; when students's scores on sequentially designed tests such as the current computerized version of the GRE are compared; and when scores in complex educational surveys with incomplete block designs for administering test questions are analyzed. How do we compare performances of different students based on different stimuli?

Equating is clearly an experimental design question. In the examples above, the experimenter has enough control over the design to make the missingness plausibly ignorable (in the sense of Rubin, 1987; summarized in Gelman, Carlin, Stern and Rubin, 1995, Chapter 7) or modelable (e.g., Chang and Ying, 1996). When Val distinguishes his X_i from the latent traits usually defined in item response models, calling it the "mean classroom achievement of the ith student, *in classes selected by student* i," the point is that there is a missingness process, in contrast to most general treatments of item response models, and it is not ignorable: clearly, students' self-selection process as well as the grades is informative for student rankings. A model such as Val's which treats this missing data mechanism as ignorable or missing at random is vulnerable to severe biases in estimation of the X_i, as illustrated recently by Bradlow and Thomas (1997). Mislevy and Wu (1996) provide general conditions needed for ignorability in various inferences from educational testing data, and Wang, Wainer and Thissen (1995) explore this problem in the context of equating exam scores when students are allowed to choose among several essay questions to answer.

2.2 Multidimensional Extensions of the Model

We were surprised that Val's one-component latent variable model for GPA adjustment did not provide much improvement over a simpler additive linear model tried by Larkey and Caulkin (1992). On the other hand, Val can and did expand to a multicomponent model, which is not possible with the Larkey–Caulkin approach. Such expansions are interesting substantively (we even found ourselves wondering if a single composite of "math" and "verbal" ability would adequately capture performance across the spectrum of undergraduate courses our statistics departments offer) as well as pragmatically; and we are happy to see that Val is exploring further possibilities along these lines as well. Stricker et al. (1994) considered a related multicomponent model for GPA.

On a pragmatic level the multicomponent model provided a meaningful improvement in the correlation of adjusted GPA ranks with other predictors. So it was disappointing to us that such extensions are apparently being dismissed in the initial proposal for the adjusted GPA at Duke as being too hard for clients and users to understand. Though Val writes that a multicomponent model may be introduced later, after a certain "comfort level" with

the one-component model has been achieved, we are reminded of the maxim "nothing is as permanent as an interim solution"; and we hope that the one-component model does not become so enshrined in administrative habit, registrar's computer programs and the like that the opportunity for later improvement is lost.

There is a larger issue concerning education of clients and collaborators. If the multicomponent model does a better job and is substantively sensible, then this is the model that should be promoted. To report merely a total score (unadjusted GPA) or unidimensional domain score (one-component adjusted GPA) is to imply to clients and users that there really is a total ordering of students that we are trying to estimate. However much clients and collaborators may dislike partial orders, if that is the reality, then we must find a way to express it in such a way that they can use the information.

A compromise that has been explored in the educational statistics literature is to work with a "dominant" unidimensional trait, but be very explicit that ignored "minor" traits introduce dependence into the likelihood that typically increases uncertainty and biases in inferences for the latent trait (Stout, 1990; Junker, 1991; Junker and Stout, 1994). Val's weighted-average achievement index based on the two-component model is a model-based version of this compromise; perhaps this would provide an opportunity to make room for the multicomponent model as it becomes more well understood.

2.3 Computation

We thought the data augmentation approach and corresponding prior for collapsing grade categories, discussed at the end of Section 2, was clever and interesting. The related discussion in Section 1.2 indicates the delicacy of the problem; for example, Bradlow (1994) shows that changing the grade category cutoff you fix can have dramatic effects on the MCMC convergence rates.

We want to raise two concerns about this prior choice, however. First, while suppressing unobserved grade categories works well for fitting observed data for one cohort of students, it may be necessary, for the related problem of predicting future grades, to allow for the occurrence of such grade categories by putting nonnegligible prior probabilities on them. Second, it appears that classes in which all students are assigned the same grade will be ignored by the model; there is little comparative information from which to assess the class's difficulty. It may be possible to use historical grade data to supply prior information on the difficulty of such classes, as long as the problem of equating across years can be overcome.

Turning to a different matter, when Val shows in Section 4.1 that the achievement index gives the correct ranking in the Larkey–Caulkin data, we would like to have seen posterior probabilities for all rankings. The beauty of obtaining Monte Carlo posterior samples is that simple counts can give interesting insights: was the "reverse" order—or some other order—a close competitor in posterior probability, for example?

2.4 Poor EB Performance

We were troubled by the poor performance of the empirical Bayes (EB) model, at least as predicted by SAT and high school GPA. If anything the EB model should be guilty of "capitalization on chance," that is, overfitting in some direction. Val's argument concerning the inability of the EB model to capture the negative relationship between mean grade and class size was not entirely convincing, but it made us realize that the EB model he is considering is more constraining, not less, than the fully Bayesian baseline model: as Val points out in specifying the priors for the baseline model, it is important to design the model so that unassigned grade categories are given negligible posterior probability. The structure of the baseline model developed in Section 2 elegantly guarantees this, but it appears that the structure for the EB model changes so that this is no longer guaranteed. Thus it is not just EB estimation of hyperparameters but also a change in the structure of the model that distinguishes the EB model from the baseline model, and the particular structural change appears to be one that Val has identified as inimical to good inferential performance.

3. CONCLUSIONS

Despite our friendly criticisms—the most important of which are shared by unadjusted GPA—we think Val's implementation of a GRM-adjusted GPA is a real improvement over unadjusted GPA, because it more fairly compares students's performances.

Val's work is an excellent example of the application of modern statistical methods to educational research. We find this domain, educational statistics, an exciting area to think about. Educational measurement data, whether from grades or tests, provide rich data structures where individual level covariates and heterogeneity are likely to allow for interesting statistical models. Typically educational data is continuously collected on large subject pools,

which affords the possibility of predictive and internal validity studies not available in other types of research. Perusal of the papers and journals referenced in this discussion, and in Val's article, will lead the reader to a vast number of interesting statistical problems, easily competitive with the best problems in biostatistics, for example. Also, as shown by the large influence at Duke and in the popular press (e.g., Pedersen, 1997) that this work has already had, statistical research with educational data and is likely to have great real-life impact.

ACKNOWLEDGMENTS

This research supported by NSF Grant DMS-94-04438 (Junker) and a grant from the Educational Testing Service, Statistics Research Group (Bradlow).

Rejoinder

Valen E. Johnson

I would like to begin by thanking all discussants for the many insightful comments received. Larkey and Young have provided an excellent overview of the problems associated with undergraduate grading, and I agree with essentially all of their remarks. As a postscript to Larkey's comments concerning the status of the reform proposal at Duke, I would add that since receiving his discussion, the Committee on Grades has been officially reconstituted and may be rechartered as a provostial committee early next fall. In addition, new motions are expected before our Arts and Sciences Council early next semester, and these new motions will likely include provisions to collect the type of data required to substantiate the problems outlined by both White and Larkey.

In regard to the penetrating discussion of Junker and Bradlow, I have only a few technical observations to make.

The relationship of the achievement index to more standard latent trait models is a subtle one. Perhaps the best way to connect the two concepts is to suppose that a high- or infinite-dimensional trait vector is associated with every student and that a student's expected performance in each class is determined by some weighting of these traits. The performance of each student in all classes—whether observed or not—can then be summarized by a weighted trait vector with dimension equal to the number of classes. In standard latent trait analysis, one might attempt to estimate every component of the underlying student trait vector from the partially observed vector of weighted classroom traits.

In contrast, the achievement index represents the mean weighted classroom trait for those classes actually taken by a student. In estimating the achievement index, we are concerned only in estimating the average weighted trait for these classes. We are not interested in estimating the weighted classroom traits for classes not taken, nor are we attempting to estimate the underlying latent traits themselves. Thus, a mathematics major's achievement index represents an entirely different weighting of traits than does an art history major's. By defining the achievement index in this way, the missing data and selection problems alluded to by Junker and Bradlow are largely avoided. Of course, as more components are added to the achievement index, the components of the index become more closely related to standard latent traits.

Junker and Bradlow also comment on the effects of varying the grade-cutoffs that are fixed to MCMC convergence rates. In regard to these remarks, I think it is important to note that no grade-cutoffs are fixed in the proposed model. Indeed, this is one of several aspects of the model that makes it distinct from other models proposed in the item response and ordinal data literature (e.g., Albert and Chib, 1993; Bradlow, 1994; and Bradlow and Zaslavsky, 1996). As noted in the text, fixing any of the grade cutoffs makes the model unsuitable for analyzing undergraduate grade data, and I suspect the same is true for many other multirater ordinal datasets as well.

Finally, Junker and Bradlow's comments regarding EB models are well taken. As White notes, the model was criticized at Duke for not "counting" independent study courses or small seminar courses in which only one grade was assigned (90% of independent study grades at Duke are either an A+, A−, or A; the median grade in classes containing fewer than five students is an A). In response to

this criticism, the model may be revised so that EB priors are used to model grade-cutoffs in classes with fewer than, say, 10 students. In order to avoid the difficulties encountered with the EB priors that were used in the article, the hyperparameter values for these priors will be fitted hierarchically by instructor, shrinking toward departmental means. This strategy will eliminate the adverse effects caused by differences in grading patterns associated with class size, instructor and department, and will hopefully improve the accuracy of the model by incorporating what information is available from small classes into the estimation of the achievement indices.

ADDITIONAL REFERENCES

ALGINA, J. (1992). Special issue: the National Assessment of Educational Progress (Editor's note). *Journal of Educational Measurement* **29** 93–94.

BEJAR, I. I. and BLEW, E. O. (1981). Grade inflation and the validity of the Scholastic Aptitude Test. *American Educational Research Journal* **18** 143–156.

BRADLOW, E. T. and THOMAS, N. (1997). Item response theory models applied to data allowing examinee choice. *Journal of Educational and Behavioral Statistics*. To appear.

CAULKINS, J. P., BARNETT, A., LARKEY, P. D., YUAN, Y. and GORANSON, J. (1993). The on-time machines: some analyses of airline punctuality. *Oper. Res.*

CHANG, H.-H. and YING, Z. (1996). Nonlinear sequential designs for logistic item response theory models with applications to computerized adaptive tests. *Ann. Statist.* To appear.

FETTER, J. H. (1995). Questions and admissions: reflections on 100,000 admissions decisions at Stanford. Stanford Univ. Press.

GELMAN, A., CARLIN, J. B., STERN, H. S. and RUBIN, D. B. (1995). *Bayesian Data Analysis.* Chapman and Hall USA, New York.

GOSE, B. (1997). *The Chronicle of Higher Education*, March 21, A53.

HEMKER, B. T., SIJTSMA, K., MOLENAAR, I. W. and JUNKER, B. W. (1997). Stochastic ordering using the latent trait and the sum score in polytomous IRT models. *Psychometrika.* To appear.

HOLLAND, P. W. and RUBIN, D. B., eds. (1982). *Test Equating.* Academic Press, New York.

HORNBLOWER, M. (1997). Learning to earn. *Time* **149** (February 24) 34.

JOHNSON, E. G., MISLEVY, R. J. and THOMAS, N. (1994). Theoretical background and philosophy of NAEP scaling procedures. In *Technical Report of the NAEP 1992 Trial State Assessment Program in Reading* (E. G. Johnson, J. Mazzeo and D. L. Kline, eds.) Chapter 8, 133–146. Office of Educational Research and Improvement, U.S. Dept. Education, Washington, D.C.

JUNKER, B. W. (1991). Essential independence and likelihood-based ability estimation for polytomous items. *Psychometrika* **56** 255–278.

JUNKER, B. W. and STOUT, W. F. (1994). Robustness of ability estimation when multiple traits are present with one trait dominant. In *Modern Theories of Measurement: Problems and Issues* (D. Laveault, B. D. Zumbo, M. E. Gessaroli and M. W. Boss, eds.) Chap. 2. Univ. Ottawa.

LAMBERT, C. (1993). Desperately seeking summa. *Harvard Magazine* **95** (May–June) 36–40.

MISLEVY, R. J. and WU, P. K. (1996). Missing Responses and IRT ability estimation: omits, choice, time limits, and adaptive testing. Technical Report RR-96-30-ONR, Educational Testing Service, Princeton, NJ.

NATIONAL CENTER FOR EDUCATION STATISTICS (1995). *Digest of Education Statistics 1995*. U.S. Dept. Education, Washington, D.C.

PATZ, R. J. (1996). Markov chain Monte Carlo methods for item response theory models with applications for the National Assessment of Educational Progress. Ph.D. dissertation, Carnegie Mellon Univ.

PEDERSEN, D. (1997). When an A is average. *Newsweek* March 3, 64.

RUBIN, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys.* Wiley, New York.

SAMEJIMA, F. (1997). Graded response model. In *Handbook of Modern Item Response Theory* (W. J. van der Linden and R. K. Hambleton, eds.) 85–100. Springer, New York.

STOUT, W. F. (1990). A new item response theory modeling approach with applications to unidimensionality assessment and ability estimation. *Psychometrika* **55** 293–325.

STRICKER, L. J., ROCK, D. A., BURTON, N. W., MURAKI, E. and JIRELE, T. J. (1994). Adjusting college grade point average criteria for variations in grading standards: a comparison of methods. *Journal of Applied Psychology* **79** 178–183.

ZWICK, R. (1992). Special issue on the National Assessment of Educational Progress. *Journal of Educational Statistics* **17** 93–94.