



---

## An Alternative to Traditional GPA for Evaluating Student Performance

Author(s): Valen E. Johnson

Source: *Statistical Science*, Vol. 12, No. 4 (Nov., 1997), pp. 251-269

Published by: Institute of Mathematical Statistics

Stable URL: <https://www.jstor.org/stable/2246210>

Accessed: 28-02-2020 03:32 UTC

### REFERENCES

Linked references are available on JSTOR for this article:

[https://www.jstor.org/stable/2246210?seq=1&cid=pdf-reference#references\\_tab\\_contents](https://www.jstor.org/stable/2246210?seq=1&cid=pdf-reference#references_tab_contents)

You may need to log in to JSTOR to access the linked references.

---

JSTOR is a not-for-profit service that helps scholars, researchers, and students discover, use, and build upon a wide range of content in a trusted digital archive. We use information technology and tools to increase productivity and facilitate new forms of scholarship. For more information about JSTOR, please contact [support@jstor.org](mailto:support@jstor.org).

Your use of the JSTOR archive indicates your acceptance of the Terms & Conditions of Use, available at <https://about.jstor.org/terms>



*Institute of Mathematical Statistics* is collaborating with JSTOR to digitize, preserve and extend access to *Statistical Science*

# 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, the growing 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).*

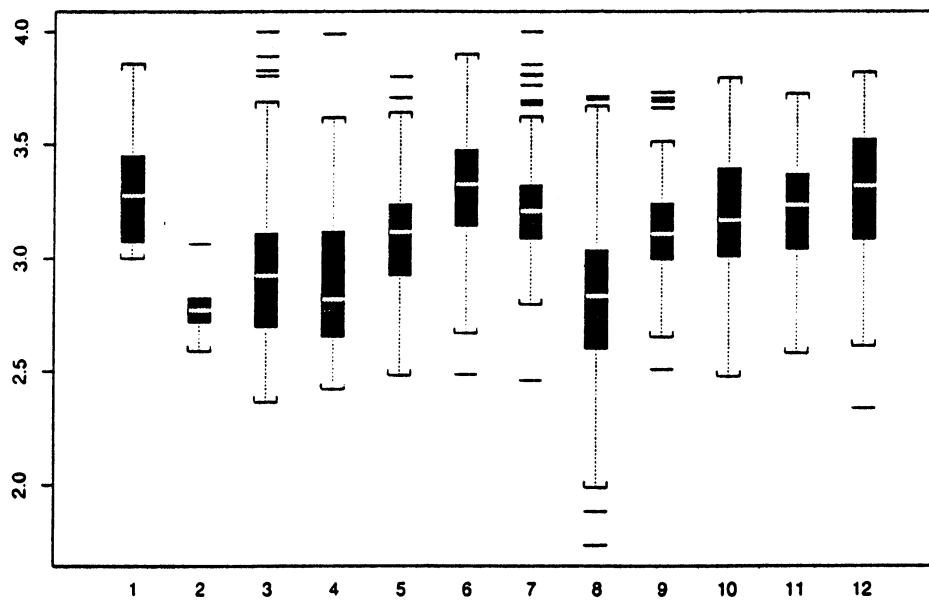


FIG. 1. Boxplots of mean classroom grades assigned in classes offered by 12 Duke University departments. The 12 departments coincide with the 12 departments studied in Goldman et al. (1974): cultural anthropology, biochemistry, biology, chemistry, economics, engineering, history, mathematics, political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

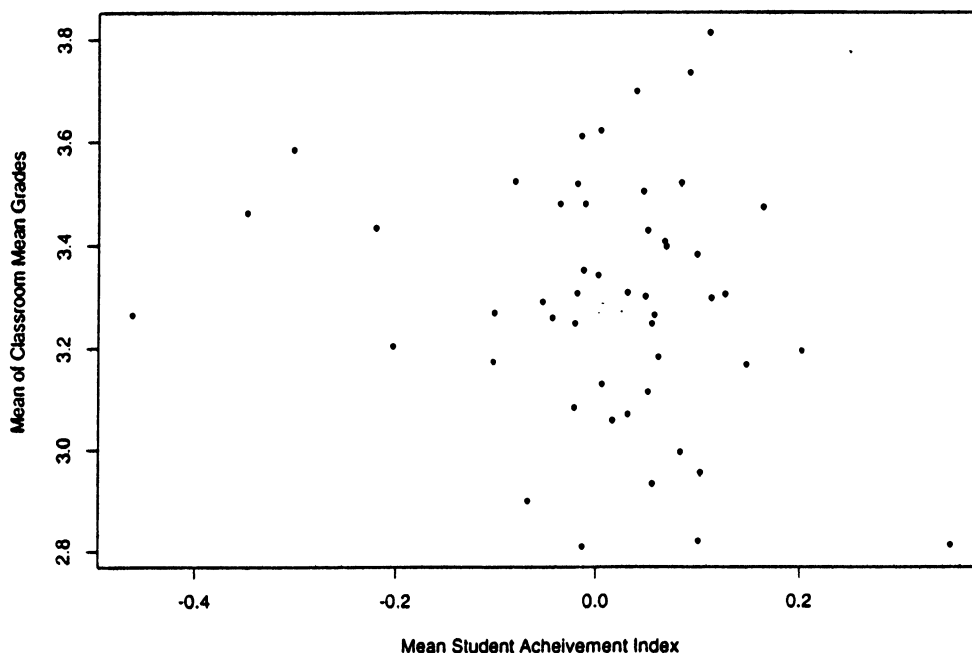


FIG. 2. Scatterplot of the mean of the mean classroom grades assigned by Duke course registration code versus the mean student achievement index of students receiving these grades. Registration codes in which fewer than 100 grades were assigned were not plotted.

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 Elliott, 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  $i$ th student's ability is denoted by  $\beta_i$ . In the terminology of IRT, let  $\eta_j$  denote the discrimination parameter of the  $j$ th class grade, and let  $\zeta_{jk}$  denote the upper grade-cutoff for grade  $k$  in class  $j$ . With this notation, the basic assumption of the GRM is

that

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

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  $\eta_j$ ,  $\beta_i$  and  $\zeta_{jk}$  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  $\pi_{ijk}$ .

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  $i$ th student in the  $j$ th class, and  $g_i$  and  $c_j$  denote the  $i$ th student's true GPA and the difficulty of the  $j$ th class, respectively, then the additive model used to estimate these adjustment factors takes the form

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

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  $i$ th student in the  $j$ th 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  $i$ th 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  $\gamma_0^j, \gamma_1^j, \dots, \gamma_K^j$ . The upper cutoff for an F in class 3 is  $\gamma_1^3$ , for a D– it is  $\gamma_2^3$  and so on up to  $\gamma_{12}^3$ , which is the upper cutoff for an A. The upper cutoff for an A+,  $\gamma_{13}^j$ , is  $\infty$ , and the lower cutoff for an F,  $\gamma_0^j$ , is  $-\infty$ .
4. Random variation in the performance of student  $i$  in class  $j$  is denoted by  $\varepsilon_{ij}$ . 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  $\varepsilon_{ij}$  is Gaussian with mean 0 and variance  $\sigma_{ij}^2$ . 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

$$(3) \quad \gamma_{k-1}^j < X_i + \varepsilon_{ij} \leq \gamma_k^j.$$

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

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

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

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

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  $\mathcal{C}_i$ .

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

$$(6) \quad X_i \sim N(0, 1),$$

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

and

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

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  $\mu$  and variance  $\tau^2$ , and  $\text{IG}(\alpha, \lambda)$

denotes an inverse gamma distribution with shape  $\alpha$  and scale  $\lambda$ . 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  $\sigma_{t(j)}^2$ .

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  $\lambda$  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  $\alpha$  and  $\lambda$  are similar to those obtained using both an empirical Bayes approach for estimating  $\alpha$  and  $\lambda$ , and those obtained by taking a uniform prior on  $\alpha$  and  $\lambda$  and treating these parameters as random quantities as well. Results from the empirical Bayes and fully Bayesian approach for estimating  $\alpha$  and  $\lambda$  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  $\gamma_k^j \leq \gamma_{k+1}^j$ ) 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  $(\iota_1^j, \dots, \iota_K^j) = \mathbf{\iota}_j$  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  $\gamma_{k-1}^j = \gamma_k^j$  whenever  $\iota_k^j = 0$ , and then taking the prior density on  $\mathbf{\iota}_j$  to be

$$(9) \quad \Pr(\mathbf{\iota}_j) = \begin{cases} \frac{\varepsilon^{S_j}}{1 - (1 - \varepsilon)^K}, & \text{if } S_j \geq 1, \\ 0, & \text{if } S_j = 0, \end{cases}$$

$$S_j = \sum_{k=1}^K \iota_k^j,$$

where  $\varepsilon \ll 1$ .

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

$$(10) \quad \Pr\left(\iota_i^j = 1 \mid \sum_{\substack{k \neq i \\ 1 \leq k < K}} \iota_k^j \geq 1\right) = \frac{\varepsilon}{1 + \varepsilon},$$

$$\Pr\left(\iota_i^j = 0 \mid \sum_{\substack{k \neq i \\ 1 \leq k < K}} \iota_k^j = 0\right) = 0.$$

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  $\iota_j$ , a non-informative (Jeffrey's) prior is assumed for the unique elements of the grade cutoffs in the  $j$ th 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

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

Here  $I = \{k: \iota_k = 1\}$ ,  $i_k$  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  $i$ th student in the  $j$ th 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

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

In (12),  $\text{Ind}(\cdot)$  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  $\pm 9.99$ . 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

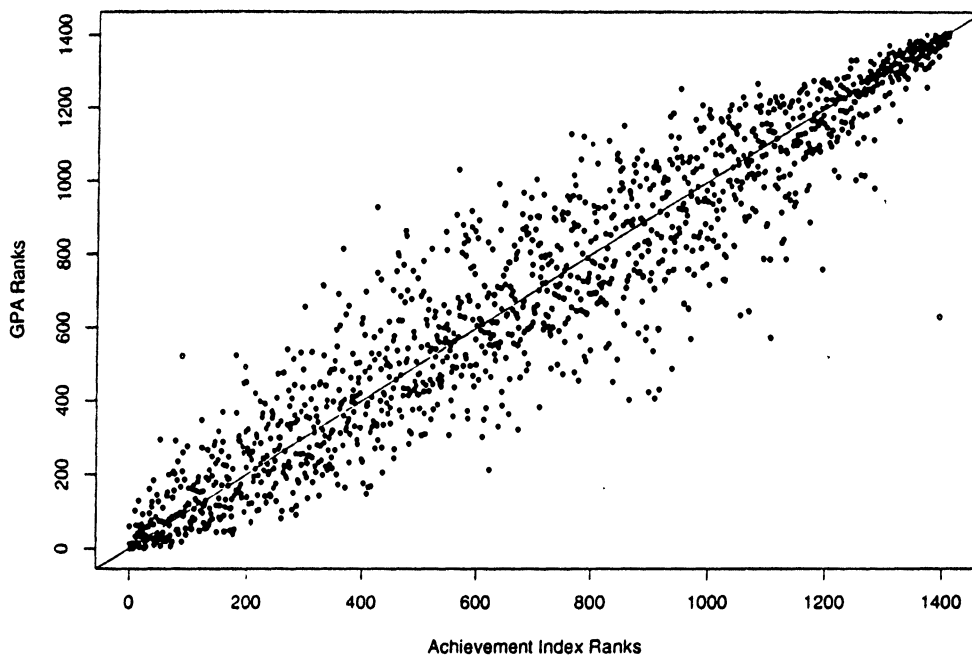


FIG. 3. Scatterplot of student ranks based on GPA versus student ranks based on achievement index for a recent class of Duke undergraduates.

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

$$(13) \quad r_{ij} = \frac{x_{ij}^* - x_i^{-j}}{\sigma_{t(j)}}.$$

Here,  $x_i^{-j}$  is the estimated posterior mean of the  $i$ th 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  $\sigma_{t(j)}^2$ , the distribution of  $x_{ij}$  is  $N(x_i, \sigma_{t(j)}^2)$ . Ignoring the dependence in the posterior distribution of  $\sigma_j^2$  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

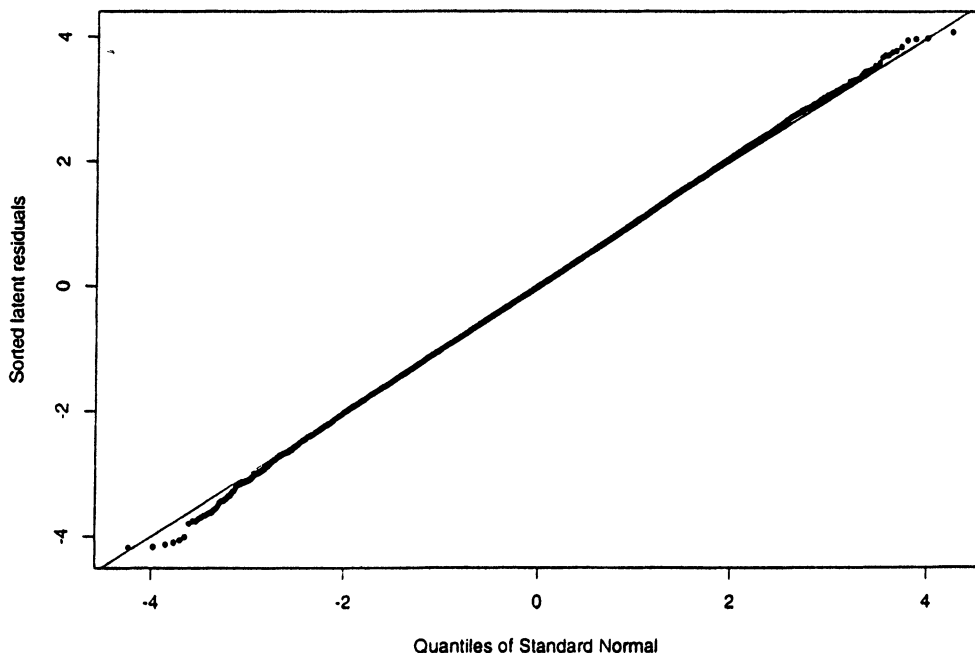


FIG. 4. Normal scores plot of sampled latent residuals.

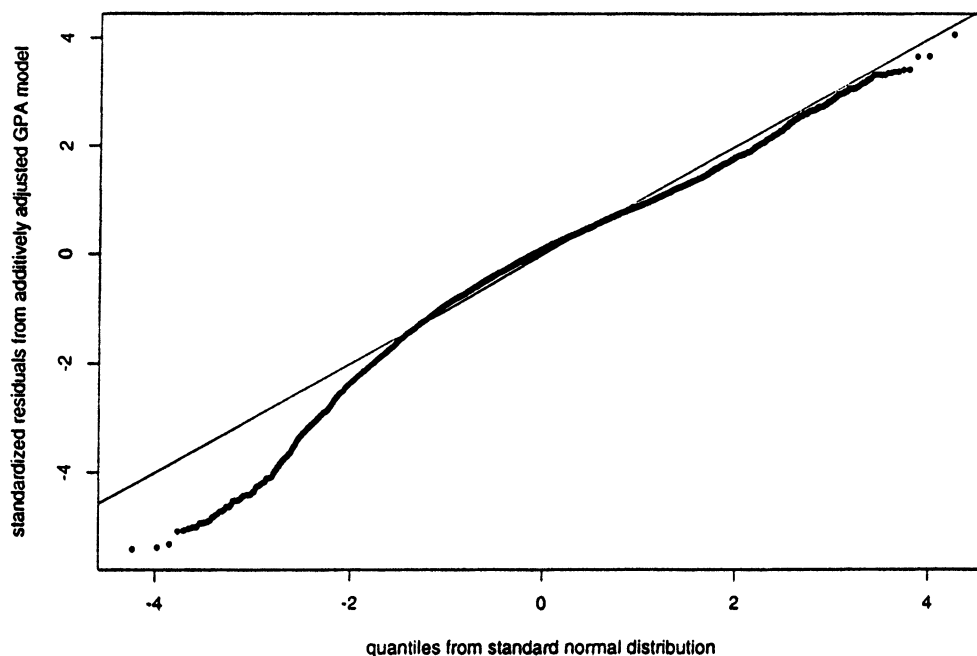


FIG. 5. Normal scores plot of sampled latent residuals.

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 hav-

ing the form

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

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  $w_d$ ,  $0 \leq w_d \leq 1$ . In the expanded model, student  $i$  is assumed to receive a grade of  $k$  in class  $j$  in department listing  $d$  whenever

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

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

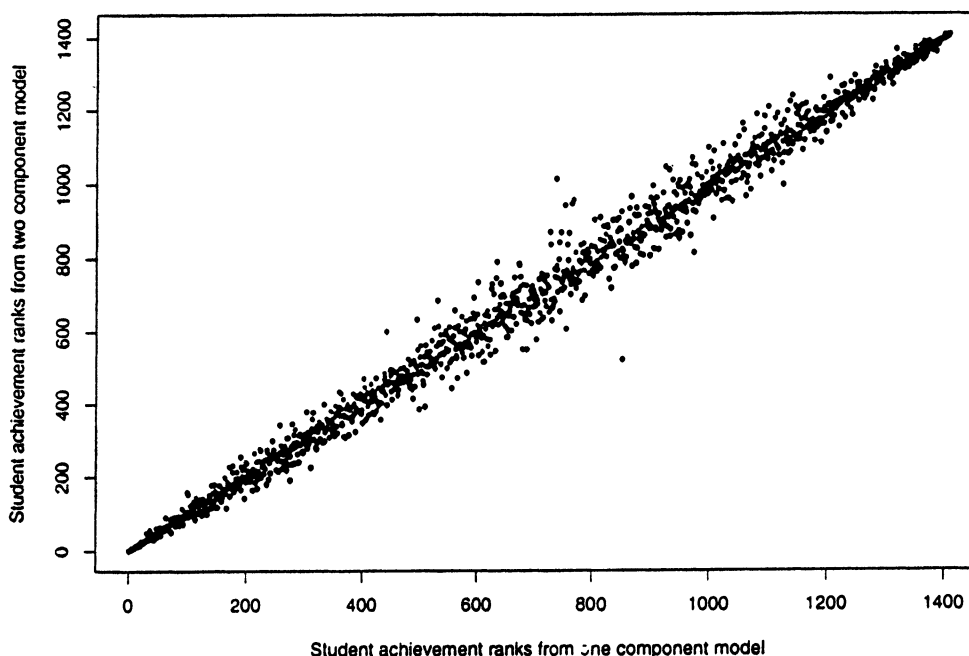


FIG. 6. Rank of student achievement indices under one- and two-component models.

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

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

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

$$(16) \quad p(\boldsymbol{\gamma}^j | \mathbf{u}) \propto \Phi(\gamma_{I_{\min}}^j / \sigma_{d(j)})^{-1/2} \cdot \left\{ \prod_{\substack{i_k \in I \\ i_k > I_{\min}}} \left( \Phi\left(\frac{\gamma_{i_k}^j}{\sigma_{d(j)}}\right) - \Phi\left(\frac{\gamma_{i_{k-1}}^j}{\sigma_{d(j)}}\right) \right) \right\}^{-1/2} \cdot \left\{ \prod_{\substack{i_k \in I \\ i_k < I_{\max}}} \frac{\phi(\gamma_{i_k}^j / \sigma_{d(j)})}{\sigma_{d(j)}} \right\}.$$

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

$$(17) \quad X_i^w = \frac{1}{|\mathcal{C}_i|} \sum_{j \in \mathcal{C}_i} (w_{d(j)} X_i^1 + (1 - w_{d(j)}) X_i^2).$$

Class ranks based on  $\{X_i^w\}$  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  $\pi_j$  denote the probability vector describing the multinomial probabilities that students in class  $j$  are assigned to different grade categories, and assuming that  $\pi_j$  is drawn from a Dirichlet distribution with parameters  $\omega = (\omega_1, \dots, \omega_K)$ . That is, assume that

$$(18) \quad p(\pi_j) \propto \prod_{k=1}^K (\pi_k^j)^{\omega_k-1},$$

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  $\omega$ , denoted here by  $\hat{\omega}$ . Transforming to the standard normal scale, the resulting empirical Bayes prior on grade-cutoffs is

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

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  $(\alpha, \lambda)$ . 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  $\alpha$  and  $\lambda$ , sample transcripts can be generated to mimic the observed values of these quantities. Implementing this moment-matching procedure yielded approximate values for  $\alpha$  and  $\lambda$  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

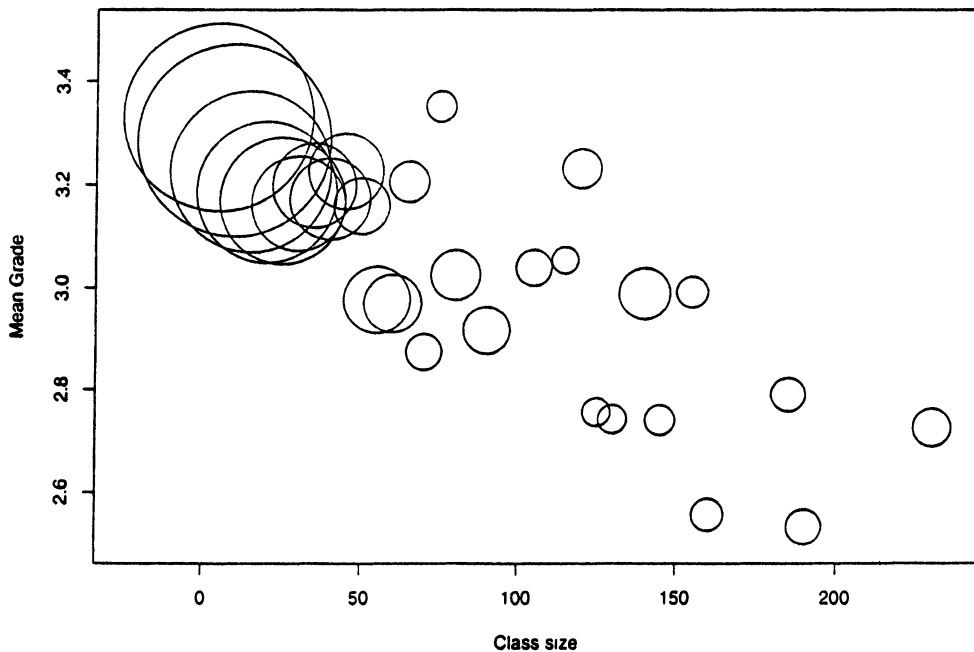


FIG. 7. Plot of mean grade assigned versus number of students in class; circles in the plot have radii proportional to the square root of the number of students used to compute the mean grade. Class sizes were grouped in units of five students. That is, the leftmost circle represents the mean classroom 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  $\alpha$  and  $\lambda$ . Unfortunately, independent uniform priors on  $\alpha$  and  $\lambda$  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  $(\alpha, \lambda)$  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  $\alpha$  and  $\lambda$ . 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  $\alpha$  and  $\lambda$  were employed. For numerical stability, grade-cutoffs were assumed to be uniformly distributed within the interval  $(-10, 10)$ .

The posterior mean of  $(\alpha, \lambda)$  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.

## 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  $\iota$  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  $\iota$ , 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

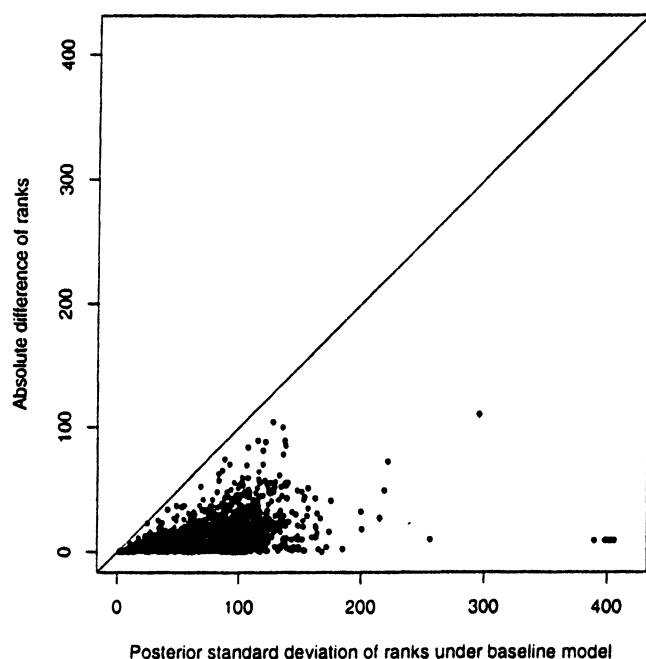


FIG. 8. *Posterior variability of AI-based class rank: this figure depicts the posterior standard deviation of the rank of each Duke student under the baseline model versus the absolute difference of the ICM ranks obtained under the baseline model and model U2. The area under the line contains those 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.*

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-