Discussion of Val Johnson's *Alternatives to GPA-based Evaluation of Student Performance*[*]

Brian W. Junker
Department of Statistics
Carnegie Mellon University
232 Baker Hall
Pittsburgh PA 15213 USA
`brian@stat.cmu.edu`
(412) 268-8873

Eric T. Bradlow
Wharton School of Business
University of Pennsylvania
Suite 1400 SH-DH
Philadelphia PA 19104-6371
`ebradlow@wharton.upenn.edu`
(215) 898-8255

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.

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 importantly 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.

[*]This research supported by NSF grant #DMS-94.04438 (Junker) and a grant from the Educational Testing Service, Statistics Research Group (Bradlow).

grades) using latent variables in a Bayesian framework: see for example Albert and Chib (1993), Bradlow (1994), 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 for example Holland and Rubin (1982). Equating problems also arise when SAT scores are compared from one year to the next; when students' 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 et. al 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 i^{th} 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 are 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 (1997) 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—not possible with the Larkey and 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.

On a pragmatic level the multi-component 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 implementation of the adjusted GPA at Duke as being too hard for clients and users to understand. Though Val writes that a multi-component 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 multi-component 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 multi-component 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 occurence of such grade categories by putting put non-negligible 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 and 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", i.e. 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' 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. And, as seen by the large influence at Duke and in the popular press (e.g. Pedersen, 1997) that this work has already enjoyed, statistical research with educational data is also likely to have great real life impact.

Additional References

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

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.

Chang, H.-H. and Ying, Z. (1996). Nonlinear sequential designs for logistic item response theory models with applications to computerized adaptive tests. Submitted for publication.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (1995). *Bayesian Data Analysis.* New York: Chapman and Hall USA.

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. In press, *Psychometrika.*

Holland, P. W. and Rubin, D. B. (Eds.) (1982). *Test equating.* New York: Academic Press.

Johnson, E. G., Mislevy, R. J., and Thomas, N. (1994). Theoretical background and philosophy of NAEP scaling procedures. Chapter 8, pp. 133–146 in Johnson, E. G., Mazzeo, J. and Kline, D. L. (1994). *Technical Report of the NAEP 1992 Trial State Assessment Program in Reading.* Washington, DC: Office of Educational Research and Improvement, U.S. Department of Education.

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. Chapter 2 in: Laveault, D., Zumbo, B.D., Gessaroli, M.E., & Boss, M.W., Eds. (1994). *Modern Theories of Measurement: Problems and Issues.* Ottawa, Canada: University of Ottawa.

Mislevy, R. J. and Wu, P. K. (1996). Missing Responses and IRT Ability Estimation: Omits, Choice, Time Limits, and Adaptive Testing. ETS Technical Report, RR-96-30-ONR.

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 University.

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

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

Samejima, F. (1997). Graded response model. pp. 85–100 in van der Linden, W. J. and Hambleton, R. K. (eds) (1997). *Handbook of modern item response theory.* New York: Springer-Verlag.

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

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.

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